COMPUTATIONAL APPROACHES FOR DEFINING THE REGULATION OF DROSOPHILA CIRCADIAN RHYTHMICITY AND PREDICTING THE ESCHERICHIA COLI PROTEOME

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Robert Steven Kuczenski

August 2008

COMPUTATIONAL APPROACHES FOR DEFINING THE REGULATION OF DROSOPHILA CIRCADIAN RHYTHMICITY AND PREDICTING THE ESCHERICHIA COLI PROTEOME

Robert Steven Kuczenski, Ph.D.

Cornell University 2008

The ever increasing ability to collect large biological datasets can quickly make analysis by traditional conceptual models intractable. Computational biology, as part of the systems biology paradigm, attempts to exploit this opportunity by providing novel hypotheses and proposing future experimentation. Here we present two computational biology efforts complementing the experimental studies of *Drosophila* circadian rhythmicity and the cell-wide regulation of protein synthesis in *Escherichia coli*. To investigate open questions of circadian regulation and communication, we developed a detailed mathematical model of *Drosophila* circadian rhythms. Using this model we investigate both the regulation of the critical PERIOD-TIMELESS negative feedback loop and the possible communication underlying the adaptation of locomotor activity to seasonal changes in day length. The model results suggest novel interactions to be tested experimentally, advancing our understanding and control of these biological rhythms. In the second part of our effort, we extended our previously developed model of protein synthesis to predict the genome-wide proteome from transcript expression data. The model results are consistent with the measured transcriptomic and proteomic changes of *E. coli* over-expressing *rhsA*. Data not well-fit by the model are consistent with post-transcriptional regulation resulting from *rhsA* over-expression. From another perspective, the spread in the experimental data is not well-described by the model results suggesting regulation based solely on codon bias may be insufficient to describe the observed global changes

in protein synthesis. Proposing a mechanism underlying the complex and nonlinear protein-mRNA relationship, the model results show significant predictive capabilities while suggesting a path to future model development. The combined results of both efforts show the potential impact of using mathematics and computers to solve complex biological problems.

BIOGRAPHICAL SKETCH

Robert Steven Kuczenski was born in Sheboygan, Wisconsin in 1980 and raised in New Holstein until 1987 when his family moved to New Brighton, Minnesota. He was educated at Holy Rosary and St. Johns elementary schools and Irondale high school. After a quarter at the University of Minnesota, Robert transferred to Iowa State University in 1999 where he obtained his B.S. in Chemical Engineering in December 2002, graduating with honors and distinction.

ACKNOWLEDGEMENTS

My academic and professional growth in graduate school has benefitted from input from my special committee, especially the indispensable advice and mentoring of Kelvin Lee and James Sethna. Their support has contributed significantly to the progress of my doctoral project.

The SloppyCell team, including Ryan Gutenkunst, Fergal Casey, Joshua Waterfall, Christopher Myers, and Jordan Atlas, have not only generously provided the computational framework on which the circadian rhythms model is built, but also offered valuable advice in numerous meetings.

I would like to thank my fiancée Jessica Anderson for her support throughout my graduate education, made all the more challenging from hundreds of miles away. I would also like to thank my many friends and classmates at Cornell University for their continued support and my new friends and colleagues at the University of Delaware who eased my group's transition to the first state.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The ever increasing ability to collect large biological datasets can quickly make analysis by traditional conceptual models intractable. Additionally, as the understanding of complex biological systems becomes more detailed, the intuitive interpretation of hypothesis-testing experimentation may become nontrivial. The fields of systems biology and computational biology have risen to meet this need by employing mathematical models of varying complexity and composition. The goal of these approaches is to accelerate the understanding of the underlying biological system by presenting novel hypotheses to be experimentally tested. Using this philosophy, mathematical models of two projects are presented here with the similar goal of providing novel biological insights. In the first project, the daily rhythms of *Drosophila* are investigated suggesting not only novel control of individually oscillating neurons but also their connectivity giving rise to emergent properties and complex behaviors. In the second project, we extend our previous model of protein synthesis to a cell-wide model. This model is used to investigate the complex and nonlinear relationship of transcriptome-proteome changes while offering insights into possible future model developments. The remainder of this Introduction details the background and motivation of these two projects. A summary of the results and impact of these projects along with proposed future work is presented in Chapter 7.

1.1 Circadian Rhythmicity

Circadian rhythmicity is a ubiquitous daily cycling found across all of kingdoms of life, from the ocean's cyanobacteria to humans and beyond. Since isolating the first muta-

tions affecting the biological clock nearly forty years ago [42], a vast experimental effort has elucidated many of the molecular cogs constituting this biological clock. Recent experimental efforts have begun to focus beyond these clock pieces to their interaction and regulation, responsible not only for daily oscillations but other emergent system properties. The similarities between the fly and mammalian clocks in addition to numerous experimental benefits including large numbers of offspring, short lifespans, and the availability of genetic tools, have made the fruit fly *Drosophila* one of the model organisms for the extensive experimental efforts studying circadian rhythmicity.

Many of the central proteins involved in producing circadian rhythms have been elucidated (reviewed in [86]). The interaction of these proteins and the underlying clock regulation are less well understood. The results of Meyer and coworkers [62] highlighted this knowledge gap when they discovered the near switch-like dissociation of two central clock proteins, PERIOD and TIMELESS prior to entering the nucleus. The timing of nuclear entry is a major regulatory point in the clock because these proteins, PERIOD in particular, complete the negative feedback loop underlying rhythmicity. The switch-like dissociation of these proteins motivated our development of a mathematical model of *Drosophila* circadian rhythmicity. By isolating the important PERIOD-TIMELESS regulatory point, several possible mechanisms were proposed and tested against experimental observations. The most consistent mechanism, one based on positive feedback, was put into the full circadian network with the results tested against numerous wild-type and mutant observations, presented in Chapter 2.

While understanding individually oscillating neurons elucidates the regulation present within individual pacemakers, the emergent properties of the interacting network of pacemaker cells is thought to give rise to many of the health-related consequences including jet lag and increased cardiovascular risk [22, 34, 95]. *Drosophila*

is again a model organism to study the possible interconnectivity, where the seasonal adaptation of the fly locomotor activity to changing day lengths has been shown to be a product of the pacemaker connectivity [89] and provides a model system for studying circadian communication. Using our previously developed mathematical model of circadian rhythmicity, we extend our investigation to the interactions of a small number of inhomogeneous oscillating neurons. In Chapter 3, a systematic search of possible pacemaker communication mechanisms is used to explore the interactions coupling the circadian pacemaker network underlying seasonal adaptation.

1.2 Bacterial Proteome

The central dogma of molecular biology preaches that DNA (the genome) is transcribed to produce RNA (the transcriptome) which is translated into protein (the proteome). While proteomic measurement techniques have advanced significantly, the coverage and cost of such measurements still significantly lags transcriptomic measurements. As a result transcript changes are commonly assumed to reflect changes in the proteome, an assumption which appears to contradict the complex and nonlinear changes in mRNA and protein observed across several organisms including *Escherichia coli* [48, 49], yeast [31, 38, 71], and the human liver [4].

Motivated by experimental observations of both the transcriptome [2] and proteome [1] by our group, we extend our previous mathematical model of protein synthesis [79] from the prediction of individual genes to cell-wide proteomic predictions. Presented in Chapter 4, the model results are found to be consistent with experimental observations while highlighting possible post-transcriptional regulation not captured in the model description. Further analysis of the model representation examines whether the model

results predict the expected heterogeneity in protein synthesis rate changes, presented in Chapter 5. Finally, the current mathematical model used in the previous two chapters may be solved using a mean-field approximation, which has an error propagation issue affecting the calculation accuracy [35, 79], or stochastically, which requires significant computational resources. The development of a new exact deterministic model of protein synthesis is presented in Chapter 6 to address these challenges, and although having promising potential, significant computational challenges remain.

CHAPTER 2

PERIOD-TIMELESS INTERVAL TIMER MAY REQUIRE AN ADDITIONAL FEEDBACK LOOP*

In this study we present a detailed, mechanism-based mathematical framework of *Drosophila* circadian rhythms. This framework facilitates a more systematic approach to understanding circadian rhythms using a comprehensive representation of the network underlying this phenomenon. The possible mechanisms underlying the cytoplasmic "interval timer" created by PERIOD-TIMELESS association are investigated, suggesting a novel positive feedback regulatory structure. Incorporation of this additional feedback into a full circadian model produced results that are consistent with previous experimental observations of wild-type protein profiles and numerous mutant phenotypes.

2.1 Introduction

Circadian rhythmicity is the product of a robust [27], free-running, temperature-compensated [99], and adaptable [16, 53] biological clock found in diverse organisms ranging from bacteria to humans. The model organism *Drosophila* is commonly used to study this phenomenon due to the relative ease of experimentation and the similarities to the mammalian circadian clock (reviewed in [86, 96]). The *Drosophila* circadian clock is composed of two interlocking feedback loops, shown in Figure 2.1. The first loop is composed of the negative feedback of *period* (*per*) and *timeless* (*tim*), shown in red, which down-regulate their own expression by inhibiting the CLOCK-CYCLE (CLK-CYC) transcription factor. DOUBLE-TIME (DBT) binds to and phosphorylates

*Article reprinted with permission from *PLoS Computational Biology*: Robert S. Kuczenski, Kevin C. Hong, Jordi García-Ojalvo, and Kelvin H. Lee, *PLoS Computational Biology* **3**, e154 (2007).

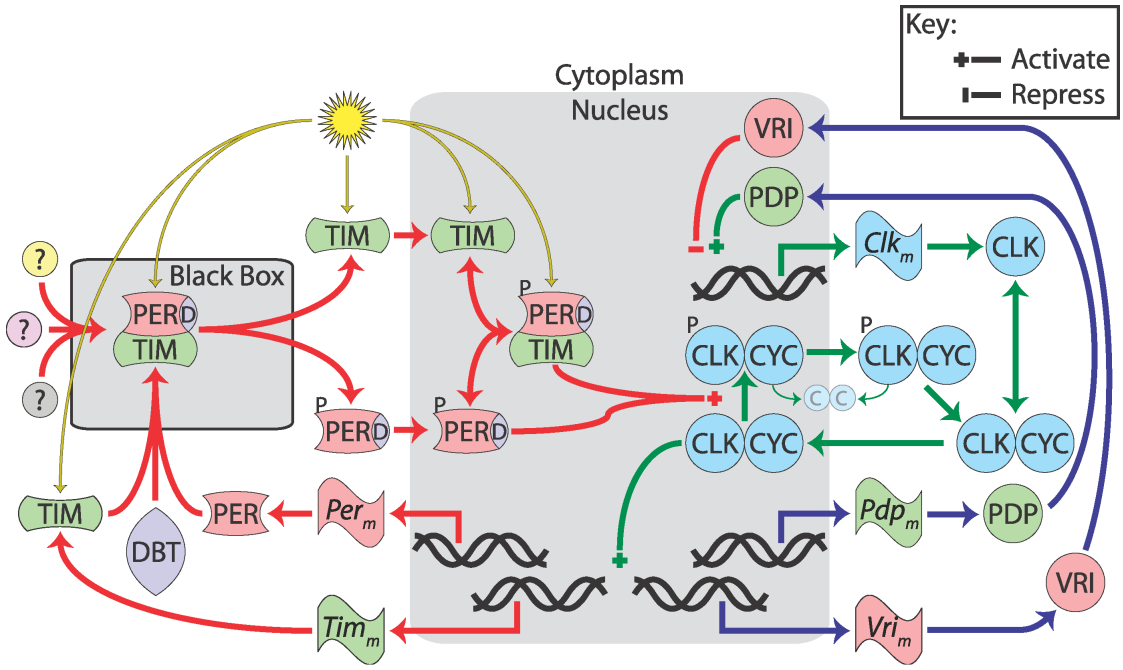

Figure 2.1: A detailed framework of circadian rhythms in *Drosophila*.

PER, which dimerizes with TIM before localizing to the nucleus via an uncharacterized mechanism. Circadian rhythms are entrained by light through an increased degradation of TIM protein, shown in yellow. In the second loop, shown in blue, the expression of *clk* is regulated by *vrille* (*vri*) and *PAR domain protein 1 isoform* ϵ (*pdp*). Both *vri* and *pdp* expression are activated by CLK-CYC. VRI represses the expression of *clk*, creating a negative feedback loop, whereas PDP creates a positive feedback loop through activating *clk* expression. Incorporating detail on interlocked feedback loops, recently shown to increase the stability and robustness of oscillations [15, 88], may be important to accurately capture the network behavior.

Several mathematical models have been created to better characterize the network underlying circadian rhythmicity in *Drosophila* (e.g., [50, 76, 82, 83, 91, 92]). These initial studies provided important insights into the molecular mechanisms of circadian rhythms and the ability to produce robust 24-hour oscillations. However, recent experi-

mental observations have created a more detailed view of network interactions, including new critical aspects that are not described by previous models. It is thus necessary to establish whether a mathematical model of the current consensus network would provide robust oscillations.

The nuclear localization of PER and TIM and the subsequent repression of CLK activity have been two particularly active areas of experimental research. The necessity of TIM for PER nuclear localization has been long established [94] and was assumed to occur through the nuclear transport of PER-TIM dimers. In contrast to this mechanism, recent experimental observations now suggest that PER and TIM localize to the nucleus in a primarily independent mechanism [5, 14, 18, 62, 65, 74]. Additionally, while TIM is required for PER nuclear localization via a cytoplasmic "interval timer" the mechanism controlling this timer is independent of both TIM and PER concentration [62]. Thus, the way in which TIM affects PER nuclear localization is an open question. Once in the nucleus, PER (and to a much lesser extent TIM) repress CLK activity, recently observed to occur via PER-mediated phosphorylation of CLK by DBT [41, 97]. These studies also provided evidence that total levels of CLK remain nearly constant [41, 97], in contrast to previously observed CLK oscillations [6, 7, 47]. It remains unclear whether constant total CLK concentration can coincide with stable oscillations in this new network.

To address these questions, we first study the possible mechanisms underlying the PER/TIM cytoplasmic interaction in *Drosophila* S2 cell culture using mathematical models of this isolated (arrhythmic) network. Using the most likely candidate mechanism, one based on positive feedback, we created a detailed mathematical model of the wild-type *Drosophila* circadian network. This model incorporates post-translational modifications to the PER and CLK proteins in addition to including both interlocked feedback loops, without the use of explicit time delays. The results of this model are

consistent with wild-type and mutant experimental observations, provide insight into recent network revisions, and suggest possible experimental directions to explore.

2.2 Results

2.2.1 Isolation of *per/tim* Feedback Loop

To investigate the six-hour delay created by the cytoplasmic interval timer observed in S2 cell culture by Meyer et al. [62], the dynamics of the *per*/*tim* loop were isolated and studied independently of the remaining circadian gene network to mimic the environment within *Drosophila* S2 cells. The interactions constituting the three mathematical models studied are shown in Figure 2.2. All models of the isolated *per*/*tim* loop include PER-TIM dimers in the cytoplasm that dissociate immediately prior to nuclear localization and re-association, but differ in the mechanism controlling the timing of this dissociation.

The mass action model is the simplest isolated model and is based on the commonly accepted *per*/*tim* interactions shown in Figure 2.2A. In this model, PER-TIM dimers simply dissociate prior to independent nuclear transport. The dynamics of this model, shown as dotted lines in Figure 2.3A, were able to produce nuclear localization of PER six hours after inducing expression, but did not accurately capture the switch-like dissociation of PER-TIM observed experimentally [62].

Next, a model based on the sequential modification of PER-TIM dimers, termed the serial model, was created as shown in Figure 2.2B. The serial mechanism may represent the sequential phosphorylation of PER and/or TIM. To simplify the mathematics of

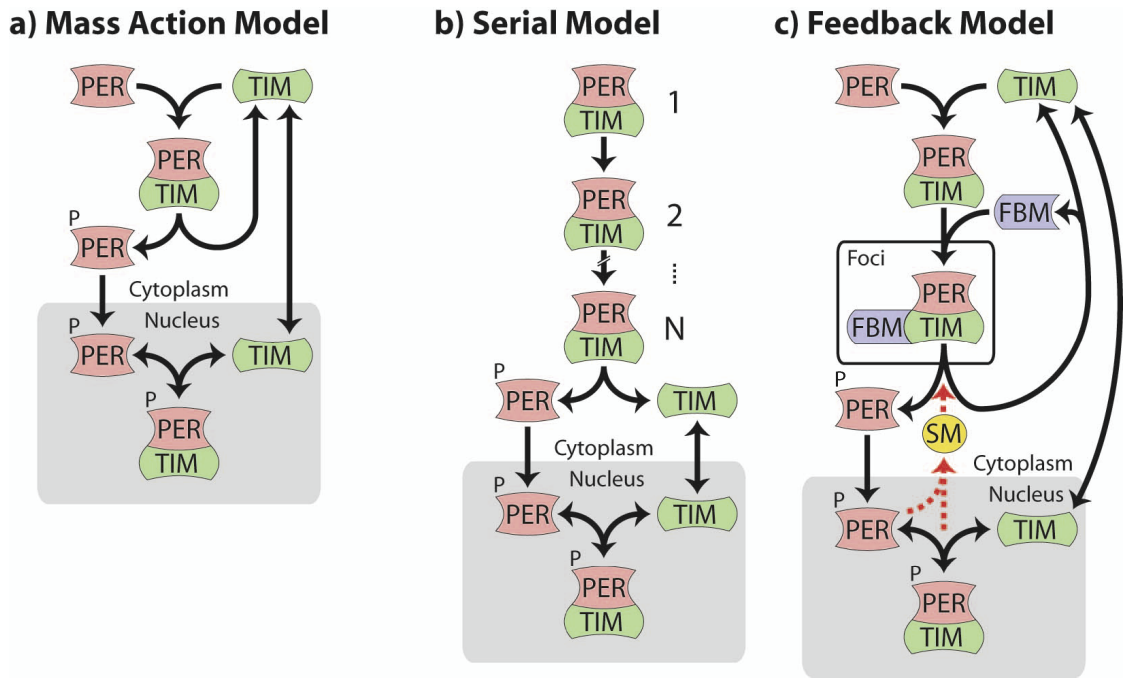

Figure 2.2: Models of the isolated *per*/*tim* loop. (A) The simple mass action kinetics model. (B) The serial model is based on a series of intermediate (possibly phosphorylated) PER-TIM states. (C) The feedback model proposes a new role for PER providing positive feedback on the dissociation of cytoplasmic PER-TIM complexes.

this model (see Materials and Methods), PER-TIM dimers were assumed to be initially associated before the proceeding series of modifications after which nuclear localization occurs. Interestingly, this model required hundreds of intermediates to produce a stable five-hour association followed by a precipitous dissociation, as shown in Figure 2.3B.

Finally, a model based on positive feedback (previously suggested to increase clock accuracy via the PDP loop of the full circadian network [26, 33]) was created as shown in Figure 2.2C. Consistent with experimental observations [62], this model explicitly represents the cytoplasmic association of PER-TIM dimers and the subsequent localization of these dimers into discrete foci. Within the foci, a background level of activity creates a low amount of dissociation and PER nuclear localization. A nuclear-generated

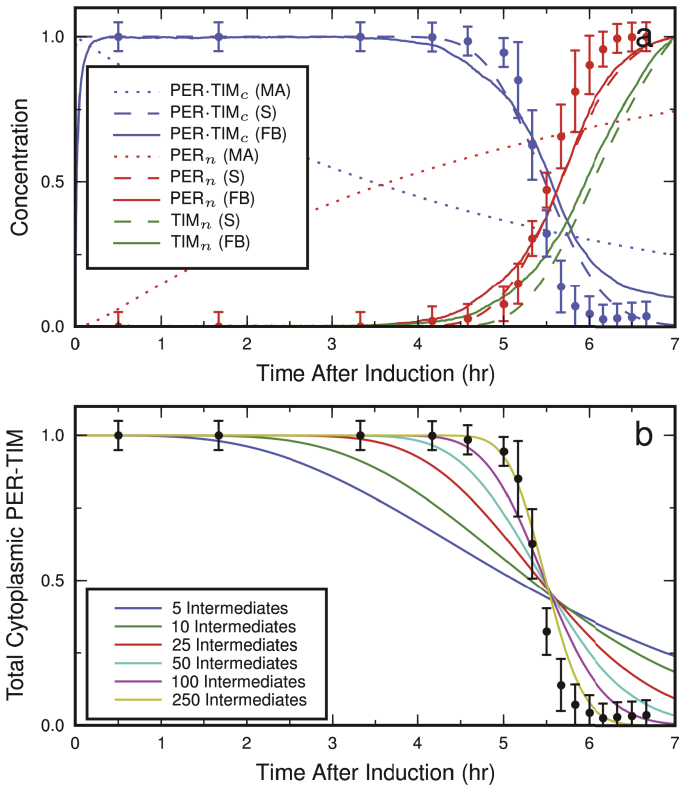

Figure 2.3: Results of the isolated *per*/*tim* loop models versus experimental data adapted from Meyer et al. [62] (points with error bars). (A) Five hours after induction, the isolated *per*/*tim* loop from the simple mass action (MA) model results show PER is nuclear but without the precipitous change in PER-TIM stability and PER localization. Both the serial (S) (100 intermediates) and the feedback (FB) model results are consistent with experimental observations, showing a precipitous dissociation or PER-TIM followed by the nuclear accumulation of PER and TIM. (B) The serial model requires hundreds of intermediate states to be consistent with experimental observations. Model results and experimental data were scaled to a maximum of 1.0. Subscripts denote cytoplamic (c) or nuclear (n) localization.

signaling molecule (SM) is then created in response to PER and is used to complete the positive feedback on the dissociation of PER-TIM in foci. This network is conceptually consistent with the observation that blocking nuclear export (and thus preventing the SM in this model from exiting the nucleus and exerting the positive feedback) delays nuclear localization [62]. The timing of PER nuclear localization in this model, shown as solid lines in Figure 2.3A, is consistent with experimental observations [62]. In addition to the feedback SM, this model incorporates another unknown component: the focus-binding mediator (FBM) molecule. The presence of this molecule at limiting concentrations creates a nuclear localization timer that is largely independent of the maximum PER and TIM concentration, as shown in Figure 2.4A.

2.2.2 Wild-Type Observations

A model of the full circadian network was created based on a simplification of the positive feedback isolated *per*/*tim* loop model, the interactions of which are shown in Figures 2.1 and 2.5. The expression of *per*, *tim*, *clk*, *vri*, and *pdp* mRNA and total protein are in excellent agreement with experimental observations, as shown in Figure 2.6 (see references therein). The model results show a period of 24.0 hours under a light-dark cycle (Figure 2.6) and 23.8 hours in constant darkness, also consistent with experimental observations.

2.2.3 Gene Dosage Affects Period Length

Our results show a *per* dosage dependence of the period length that is consistent with experimental observations [8, 81]. A continuation analysis of the maximum transcriptional

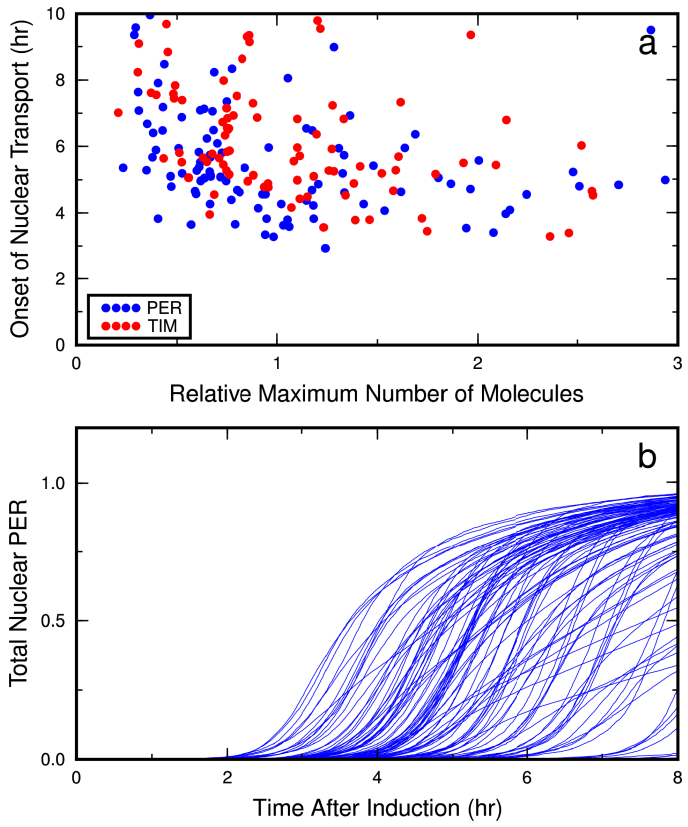

Figure 2.4: Nuclear onset in the isolated positive feedback model. (A) The onset of PER nuclear localization versus the maximum concentration of PER (blue) and TIM (red). The initial concentrations of PER and TIM were varied according to a log normal distribution (see Materials and Methods) in the stochastic implementation of the positive feedback isolated *per*/*tim* loop model. The onset of nuclear localization is largely independent of PER and TIM concentration, consistent with experimental observations [62]. (B) The ensemble of total nuclear PER trajectories used to create the blue dots in (A). Trajectories were scaled to a maximum of 1.0.

activation of *per* in the model demonstrates an inverse relation between *per* dosage and the period of circadian oscillation (black lines and points in Figure 2.7). In contrast, a continuation analysis of the maximum transcriptional activation of *tim* (gray lines and points in Figure 2.7) revealed a profile which is similar to *per* dosage and thus not very consistent with experimental observations [5, 72].

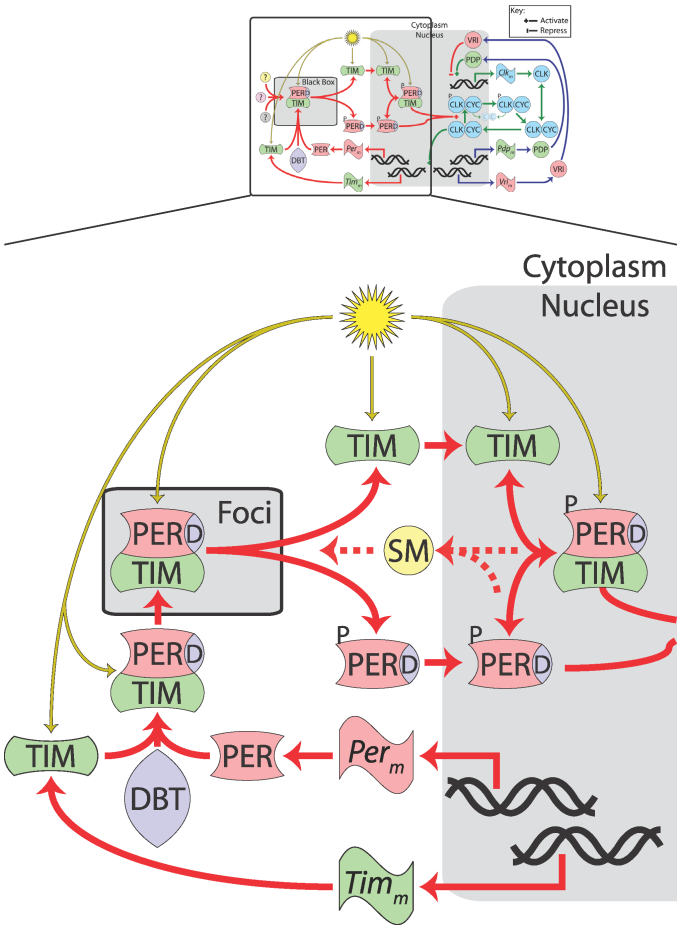

Figure 2.5: Detailed model black box. The *per*/*tim* loop of the detailed mathematical model of *Drosophila* circadian rhythmicity, a possible mechanism underlying the black box in Figure 2.1. Only one type of monomeric TIM and no monomeric DBT species were explicitly represented in the model. See Materials and Methods for a list of model equations.

2.2.4 Mutant Phenotypes

The results from the model are consistent with numerous homozygous mutant phenotypes, as shown in Table 2.1 (see references therein). These results show that arrhythmic null mutants in the *per*/*tim* feedback loop (i.e., per^{01} and tim^{01}) are unable to repress the activity of CLK-CYC resulting in constitutively high expression of unaltered *per* [84, 87], *tim* [6, 47, 64, 68, 87], *vri* [9, 17], and *pdp* [17]. The decreased PER degrada-

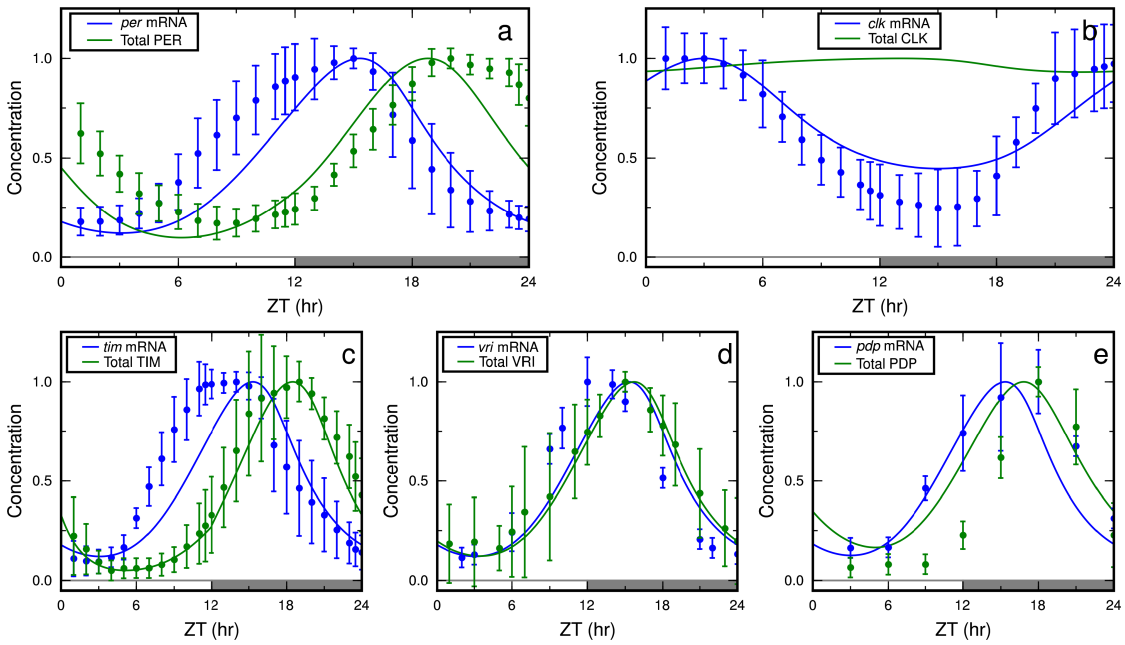

Figure 2.6: Comparison of the results with the experimental data. The expression profiles of *per* (A), *clk* (B), *tim* (C), *vri* (D), and *pdp* (E) mRNA and total protein from the detailed model results (lines) were compared with experimental observations (points with error bars generated from the average and standard deviation of published experimental data, see Table 2.5). Model results and experimental data were normalized to 1.0. Bars along the horizontal axis indicate light entrainment regime: light-on (empty) or light-off (filled).

tion in dbt^{P} and dbt^{ar} mutants resulted in the stable repression of CLK-CYC activity and the constitutively low expression of *per*, *tim*, *vri*, and *pdp* mRNA and protein [68, 73]. Similarly, when the level of active CLK-CYC is reduced by a knockout of CLK or CYC (clk^{Jrk} and cyc^{0}) or eliminating the activator of *clk* expression (pdp^{P205}), the resulting levels of *per*, *tim*, *vri*, and *pdp* mRNA and protein are constitutively low [3, 9, 17, 25, 75]. Understanding the effects of these mutants provides key insights into the roles of specific genes in the network, and reproducing their behavior provides support for the model representation.

The model accurately captures a majority of the published experimental observa-

Table 2.1: Comparison of experimental homozygous arrhythmic mutant phenotypes

Mutant	***per* mRNA**	***tim* mRNA**	***vri***	***pdp***	***clk***
			mRNA	**mRNA**	**mRNA**
	PER	**TIM**	**VRI**	**PDP**	**CLK**
per^{01}	↑ [75, 84, 87, 98]	↑ [75, 87] [84]	↑ [9, 17]	↑ [17]	↓ [7, 26]
	↓ [5, 6, 18, 20, 47, 78]	↑ [6, 47, 64, 68]	↑ [17]	↑ [17]	↑ [6, 47]
tim^{01}	↑ [84, 87] [75]	↑ [9, 75, 87] [84]	↑ [9, 17]	↑ [17]	↓ [7]
	↑ [5, 6, 18, 47, 68]	↓ [5, 6, 36, 47, 64, 78]	↑	↑ [17]	↑ [6, 47]
dbt^{P}, dbt^{ar}	↓ [68]	↓ [68]	↓	↓	↑
	↓ [68, 73]	↓ [68, 73]	↓	↓	↑
clk^{Jrk}	↓ [3]	↓ [9, 3]	↓ [17, 9]	↓ [17]	↓ [26]
	↓ [3]	↓ [3]	↓ [17]	↓ [17]	↓
cyc^{0}	↓ [75]	↓ [9, 75]	↓ [9]	↓	↓ [26]
	↓ [75]	↓ [75]	↓ [25]	↓ [17]	↓
pdp^{P205}	↓	↓ [17]	↓	↓	↓
	↓ [17]	↓	↓	↓	↓

Notes: Peak expression levels for mRNA and protein are listed relative to wild-type maxima: non-detectable to half-peak levels (↓), or greater than half-peak level (↑). References (in brackets) in green are consistent with the model results, red references are not.

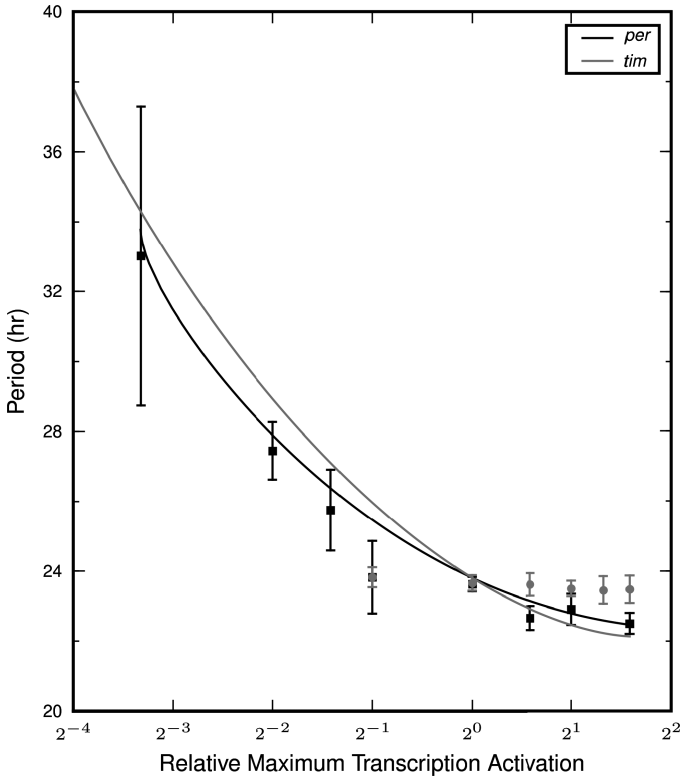

Figure 2.7: The effect of gene dosage on the period of oscillation in constant darkness. A continuation analysis of *per* dosage-dependent behavior of period under constant darkness (black line) shows an inverse relation between the maximum transcription activation of *per* and the period length, which is consistent with the experimental results (represented by black squares and error bars) [8, 81]. A continuation analysis of *tim* dosage-dependent behavior of period under constant darkness (gray line) shows a similar trend to *per* dosage, which is inconsistent with experimental observations (represented by gray circles and error bars) [5, 72].

tions. However, a number of mutant flies display behavior that is not completely consistent with the model results. For example, the low levels of *tim* mRNA in per^{01}, *per* mRNA in tim^{01}, and *tim* mRNA in tim^{01} from some publications [75, 84] conflict with model results; however, experimental results from other publications on these same species do agree with our model results [9, 75, 84, 87]. The low levels of *per* mRNA in per^{01} [75, 84, 87, 98], low levels of PER in tim^{01} [5, 6, 18, 47, 68], and high levels of *per* mRNA and protein in dbt^{P}/dbt^{ar} [68, 73] observed experimentally conflict with the

model results and experimental observations of other E-box mRNA and protein levels. The mathematical model lacks ability to produce nuclear CLK-CYC in clk^{Jrk} and cyc^{0} mutants, breaking the activation of E-box genes and producing no *clk* mRNA in contrast to experimental observations [26]. Also, the non-PER-mediated CLK phosphorylation in the model results are not able to produce low CLK levels [6, 47] without nuclear PER in the per^{01} and tim^{01} mutants.

2.3 Discussion

2.3.1 Interval Timer Control

The isolated mass action model results (dotted lines in Figure 2.3A) are not consistent with the experimental observation of stably associated PER-TIM dimers and precipitous nuclear localization [62]. The serial model results (Figure 2.3B) show that hundreds of intermediates may be required to produce this behavior. This number of intermediates is larger than the potential phosphorylation sites on PER and TIM predicted by ScanProsite (22 Casein Kinase II sites on PER, 32 sites on TIM) [23]. The progressive phosphorylation of PER and/or TIM may be observed as a change in electrophoretic mobility prior to nuclear localization in S2 cells. The positive feedback mechanism (solid lines in Figure 2.3A) is able to produce the correct delay and rapid dissociation, making it an attractive alternative to the serial model.

The FBM in the positive feedback model, for which no direct experimental evidence currently exists, is responsible for controlling the onset of nuclear localization independent of PER and TIM concentrations. Without this molecule, the onset would be well correlated to experiment-to-experiment variability in the limiting concentration of PER

and/or TIM (unpublished data). The *shaggy* (*sgg*) kinase is a potential candidate because it has been previously shown to phosphorylate TIM, affecting the nuclear localization of PER [18, 58], and also bind to cytoskeletal elements [10], a possible location of the cytoplasmic foci. A *sgg* knockout in S2 cells could be used to observe the possible disruption of PER/TIM accumulation in cytoplasmic foci, which would be consistent with this hypothesized role for *sgg*.

No obvious candidate for the SM exists in the literature. Because small molecules have been shown to cause significant structural changes in PAS domains [90], one possibility may be a small molecule binding to and elucidating a temporary conformational change in PER, allowing it to dissociate from TIM and localize to the nucleus.

The feedback model is not consistent with all the data presented by Meyer et al. [62]. The rates of nuclear localization of PER and TIM are not completely independent (unpublished data), and the conflict between rapid nuclear transport and well-controlled timing of nuclear localization results in a timing error that is double the observed seventy minutes [62]. These differences may be the result of additional regulatory structures not already identified.

2.3.2 CLK Oscillation

The full network results demonstrate that while total CLK levels do not change significantly during the course of a day, the oscillating phosphorylation of CLK can lead to significant and stable oscillations in mRNA (see Figure 2.6). These near-constant total CLK levels are generated by synchronized translation and degradation (see Figure 2.8). This result differs from previous mathematical models [82, 83, 92] which show a significant oscillation in CLK level (consistent with prior experimental observations [6, 7, 47]),

and suggest that the oscillation of CLK activity, not concentration, is necessary for circadian rhythmicity.

2.3.3 Conclusion

We find that independent transfer of PER and TIM by simple mass action kinetics is inconsistent with experimental observations, but that an additional feedback loop (or alternatively a large number of intermediate phosphorylated states) is able to produce the switch-like dissociation of cytoplasmic PER-TIM underlying the interval timer [62]. This positive feedback was introduced into a mechanistic mathematical framework for *Drosophila* circadian rhythms which demonstrates excellent agreement with experimentally observed expression profiles of circadian genes and many circadian mutants. The framework is consistent with observations of the relationship between *per* dosage and circadian periodicity. Post-translational regulation is addressed, including the effect of phosphorylation on the transcriptional activation activity of CLK. Our results also show that the nuclear translocation of the PER and TIM can occur independently while producing stable oscillations when positive feedback is employed.

2.4 Materials and Methods

With the exception of transcription activation kinetics, discussed below, all reaction kinetics are mass action. These kinetics were chosen for simplicity and because no direct experimental evidence is available suggesting other (e.g., Michaelis-Menten or Hill) kinetic forms. Unless otherwise noted, molecules use subscripts to denote mRNA (m), free cytoplasmic protein (c), focus-bound protein (f), and nuclear protein (n). Addi-

tionally, dimers are represented as $[X \cdot Y]$ and phosphorylated isoforms as $[X \cdot P]$. All models are available online [43].

2.4.1 Isolated *per/tim* Loop Models.

The simple mass action model of the isolated *per/tim* loop is represented by Equations 2.1–2.7 below.

$$\frac{\mathrm{d}\,[PER_c]}{\mathrm{dt}} = -k_{PTc} \cdot [PER_c] \cdot [TIM_c] \tag{2.1}$$

$$\frac{\mathrm{d}\,[TIM_c]}{\mathrm{dt}} = k_{MA} \cdot [PER \cdot TIM_c] - k_{PTc} \cdot [PER_c] \cdot [TIM_c] - N_T \cdot [TIM_c] + C_T \cdot [TIM_n] \tag{2.2}$$

$$\frac{\mathrm{d}\,[PER \cdot TIM_c]}{\mathrm{dt}} = k_{PTc} \cdot [PER_c] \cdot [TIM_c] - k_{MA} \cdot [PER \cdot TIM_c] \tag{2.3}$$

$$\frac{\mathrm{d}\,[PER \cdot P_c]}{\mathrm{dt}} = k_{MA} \cdot [PER \cdot TIM_c] - N_P \cdot [PER \cdot P_c] + C_P \cdot [PER \cdot P_n] \tag{2.4}$$

$$\frac{\mathrm{d}\,[PER \cdot P_n]}{\mathrm{dt}} = k_{dPTn} \cdot [PER \cdot TIM_n] - k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] + N_P \cdot [PER \cdot P_c] - C_P \cdot [PER \cdot P_n] \tag{2.5}$$

$$\frac{\mathrm{d}\,[TIM_n]}{\mathrm{dt}} = k_{dPTn} \cdot [PER \cdot TIM_n] - k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] + N_T \cdot [TIM_c] - C_T \cdot [TIM_n] \tag{2.6}$$

$$\frac{\mathrm{d}\,[PER \cdot TIM_n]}{\mathrm{dt}} = k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] - k_{dPTn} \cdot [PER \cdot TIM_n] \tag{2.7}$$

The serial model is represented by Equations 2.8–2.13 below, where N is the number of intermediates in the reaction series.

$$[PER \cdot TIM_{c,i}] = [PER \cdot TIM_{c,1}]_{t=0} \cdot \frac{(k_N \cdot t)^{i-1}}{(i-1)!} \cdot \exp(-k_N \cdot t) \quad \text{for } i = 1..N \tag{2.8}$$

$$\frac{\mathrm{d}\,[PER \cdot P_c]}{\mathrm{dt}} = k_N \cdot [PER \cdot TIM_{c,N}] - N_P \cdot [PER \cdot P_c]$$

$$+ C_P \cdot [PER \cdot P_n] \tag{2.9}$$

$$\frac{\mathrm{d}\,[TIM_c]}{\mathrm{dt}} = k_N \cdot [PER \cdot TIM_c]\, N + C_T \cdot [TIM_n] - N_T \cdot [TIM_c] \tag{2.10}$$

$$\frac{\mathrm{d}\,[PER \cdot P_n]}{\mathrm{dt}} = k_{dPTn} \cdot [PER \cdot TIM_n] - k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] + N_P \cdot [PER \cdot P_c] - C_P \cdot [PER \cdot P_n] \tag{2.11}$$

$$\frac{\mathrm{d}\,[TIM_n]}{\mathrm{dt}} = k_{dPTn} \cdot [PER \cdot TIM_n] - k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] + N_T \cdot [TIM_c] - C_T \cdot [TIM_n] \tag{2.12}$$

$$\frac{\mathrm{d}\,[PER \cdot TIM_n]}{\mathrm{dt}} = k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] - k_{dPTn} \cdot [PER \cdot TIM_n] \tag{2.13}$$

To simplify the solution of the serial model, initial concentrations of monomeric PER and TIM in the cytoplasm were eliminated by assuming that their dimerization occurred quickly. This assumption allowed the analytic solution of the last PER-TIM dimer in the series of N reactions, and greatly reduced the number of equations for large N.

The positive feedback model is represented by Equations 2.14–2.23 below. To represent the concentration effect of foci localization, a second-order term is used for slow dissociation of PER-TIM from the foci (see Equations 2.15, 2.17, and 2.18). Additionally, SM is assumed to catalyze the release of PER-TIM from the foci, and thus is not depleted by this reaction.

$$\frac{\mathrm{d}\,[PER_c]}{\mathrm{dt}} = k_{dPTc} \cdot [PER \cdot TIM_c] - k_{PTc} \cdot [PER_c] \cdot [TIM_c] \tag{2.14}$$

$$\frac{\mathrm{d}\,[TIM_c]}{\mathrm{dt}} = k_{dPTc} \cdot [PER \cdot TIM_c] - k_{PTc} \cdot [PER_c] \cdot [TIM_c] + 2 \cdot P_{PT} \cdot \left[PER \cdot TIM_f\right]^2 + k_{FB} \cdot \left[PER \cdot TIM_f\right] \cdot [SM_c] + C_T \cdot [TIM_n] - N_T \cdot [TIM_c] \tag{2.15}$$

$$\frac{\mathrm{d}\,[PER \cdot TIM_c]}{\mathrm{dt}} = k_{PTc} \cdot [PER_c] \cdot [TIM_c] - k_{dPTc} \cdot [PER \cdot TIM_c] - k_{PTf} \cdot [FBM_c] \cdot [PER \cdot TIM_c] \tag{2.16}$$

$$\frac{\mathrm{d}\left[PER \cdot TIM_f\right]}{\mathrm{dt}} = k_{PTf} \cdot [FBM_c] \cdot [PER \cdot TIM_c] - 2 \cdot P_{PT} \cdot \left[PER \cdot TIM_f\right]^2 - k_{FB} \cdot \left[PER \cdot TIM_f\right] \cdot [SM_c] \quad (2.17)$$

$$\frac{\mathrm{d}\left[PER \cdot P_c\right]}{\mathrm{dt}} = 2 \cdot P_{PT} \cdot \left[PER \cdot TIM_f\right]^2 + k_{FB} \cdot \left[PER \cdot TIM_f\right] \cdot [SM_c] + C_P \cdot [PER \cdot P_n] - N_P \cdot [PER \cdot P_c] \quad (2.18)$$

$$\frac{\mathrm{d}\left[PER \cdot P_n\right]}{\mathrm{dt}} = k_{dPTn} \cdot [PER \cdot TIM_n] - k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] + N_P \cdot [PER \cdot P_c] - C_P \cdot [PER \cdot P_n] \quad (2.19)$$

$$\frac{\mathrm{d}\left[TIM_n\right]}{\mathrm{dt}} = k_{dPTn} \cdot [PER \cdot TIM_n] - k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] + N_T \cdot [TIM_c] - C_T \cdot [TIM_n] \quad (2.20)$$

$$\frac{\mathrm{d}\left[PER \cdot TIM_n\right]}{\mathrm{dt}} = k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] - k_{dPTn} \cdot [PER \cdot TIM_n] \quad (2.21)$$

$$\frac{\mathrm{d}\left[FBM_c\right]}{\mathrm{dt}} = 2 \cdot P_{PT} \cdot \left[PER \cdot TIM_f\right]^2 + k_{FB} \cdot \left[PER \cdot TIM_f\right] \cdot [SM_c] - k_{PTf} \cdot [FBM_c] \cdot [PER \cdot TIM_c] \quad (2.22)$$

$$\frac{\mathrm{d}\left[SM_c\right]}{\mathrm{dt}} = k_{SM} \cdot ([PER \cdot P_n] + [PER \cdot TIM_n]) - D_{SM} \cdot [SM_c] \quad (2.23)$$

The initial concentrations of cytoplasmic PER and TIM in the mass action and positive feedback models and the first PER-TIM dimer in the serial model were set to the maximum concentration of PER and TIM (10,000 molecules or approximately 10^4 nM). The initial concentration of FBM in the positive feedback model was set to 5,000 molecules. All other initial concentrations in the isolated models were set to zero. Additionally, the initial conditions of PER and TIM in the positive feedback model were varied in magnitude based on a log normal distribution fit to the data presented in Figure 2.1C of [62]. See Table 2.2 for a full list of reaction rate constants for the isolated models.

2.4.2 Detailed Mathematical Model

A detailed mathematical framework of *Drosophila* circadian rhythms using the differential equations below was created based on the interactions shown in Figures 2.1 and 2.5.

$$\begin{aligned}
\frac{\mathrm{d}\,[Per_m]}{\mathrm{dt}} &= I_m \cdot \left(1.0 - \exp\left(-\frac{k_{bP}}{I_m} \cdot \frac{K_{RC} \cdot [CLK \cdot CYC_n]}{DN + K_{RC} \cdot [CLK \cdot CYC_n]}\right)\right) \\
&\quad - D_m \cdot [Per_m] \qquad (2.24)\\
\frac{\mathrm{d}\,[Tim_m]}{\mathrm{dt}} &= I_m \cdot \left(1.0 - \exp\left(-\frac{k_{bT}}{I_m} \cdot \frac{K_{RC} \cdot [CLK \cdot CYC_n]}{DN + K_{RC} \cdot [CLK \cdot CYC_n]}\right)\right) \\
&\quad - D_m \cdot [Tim_m] \qquad (2.25)\\
\frac{\mathrm{d}\,[PER_c]}{\mathrm{dt}} &= T_P \cdot [Per_m] - k_{PTc} \cdot [PER_c] \cdot [TIM_c] - D_P \cdot [PER_c] \\
&\quad + D_L \cdot Light \cdot \left([PER \cdot TIM_c] + \left[PER \cdot TIM_f\right]\right) \qquad (2.26)\\
\frac{\mathrm{d}\,[TIM_c]}{\mathrm{dt}} &= T_T \cdot [Tim_m] - k_{PTc} \cdot [PER_c] \cdot [TIM_c] - N_T \cdot [TIM_c] \\
&\quad + 2 \cdot P_{PT} \cdot \left[PER \cdot TIM_f\right]^2 + k_{FB} \cdot [SM_c] \cdot \left[PER \cdot TIM_f\right] \\
&\quad - (D_T + D_L \cdot Light) \cdot [TIM_c] \qquad (2.27)\\
\frac{\mathrm{d}\,[PER \cdot TIM_c]}{\mathrm{dt}} &= k_{PTc} \cdot [PER_c] \cdot [TIM_c] - k_{PTf} \cdot [PER \cdot TIM_c] \\
&\quad - D_L \cdot Light \cdot [PER \cdot TIM_c] \qquad (2.28)\\
\frac{\mathrm{d}\left[PER \cdot TIM_f\right]}{\mathrm{dt}} &= k_{PTf} \cdot [PER \cdot TIM_c] \\
&\quad - k_{FB} \cdot [SM_c] \cdot \left[PER \cdot TIM_f\right] - 2 \cdot P_{PT} \cdot \left[PER \cdot TIM_f\right]^2 \\
&\quad - D_L \cdot Light \cdot \left[PER \cdot TIM_f\right] \qquad (2.29)\\
\frac{\mathrm{d}\,[PER \cdot P_c]}{\mathrm{dt}} &= 2 \cdot P_{PT} \cdot \left[PER \cdot TIM_f\right]^2 + k_{FB} \cdot [SM_c] \cdot \left[PER \cdot TIM_f\right] \\
&\quad - N_P \cdot [PER \cdot P_c] - D_P \cdot [PER \cdot P_c] \qquad (2.30)\\
\frac{\mathrm{d}\,[PER \cdot P_n]}{\mathrm{dt}} &= k_{dPTn} \cdot [PER \cdot TIM_n] - k_{PTn} \cdot [PER \cdot P_n] \cdot [TIM_n] \\
&\quad + N_P \cdot [PER \cdot P_c] \\
&\quad - D_P \cdot [PER \cdot P_n] + D_L \cdot Light \cdot [PER \cdot TIM_n] \qquad (2.31)
\end{aligned}$$

$$\frac{\mathrm{d}\,[TIM_n]}{\mathrm{dt}} = k_{dPTn}\cdot[PER\cdot TIM_n] - k_{PTn}\cdot[PER\cdot P_n]\cdot[TIM_n] + N_T\cdot[TIM_c] - (D_T + D_L\cdot Light)\cdot[TIM_n] \quad (2.32)$$

$$\frac{\mathrm{d}\,[PER\cdot TIM_n]}{\mathrm{dt}} = k_{PTn}\cdot[PER\cdot P_n]\cdot[TIM_n] - k_{dPTn}\cdot[PER\cdot TIM_n] - D_L\cdot Light\cdot[PER\cdot TIM_n] \quad (2.33)$$

$$\frac{\mathrm{d}\,[SM_c]}{\mathrm{dt}} = k_{SM}\cdot([PER\cdot P_n] + [PER\cdot TIM_n]) - D_{SM}\cdot[SM_c] \quad (2.34)$$

$$\frac{\mathrm{d}\,[Vri_m]}{\mathrm{dt}} = I_m\cdot\left(1.0 - \exp\left(-\frac{k_{bV}}{I_m}\cdot\frac{K_{RC}\cdot[CLK\cdot CYC_n]}{DN + K_{RC}\cdot[CLK\cdot CYC_n]}\right)\right) - D_m\cdot[Vri_m] \quad (2.35)$$

$$\frac{\mathrm{d}\,[Pdp_m]}{\mathrm{dt}} = I_m\cdot\left(1.0 - \exp\left(-\frac{k_{bD}}{I_m}\cdot\frac{K_{RC}\cdot[CLK\cdot CYC_n]}{DN + K_{RC}\cdot[CLK\cdot CYC_n]}\right)\right) - D_m\cdot[Pdp_m] \quad (2.36)$$

$$\frac{\mathrm{d}\,[Clk_m]}{\mathrm{dt}} = I_m\cdot\left(1.0 - \exp\left(-\frac{k_{bC}}{I_m}\cdot\frac{K_{RD}\cdot[PDP_n]}{DN + K_{RV}\cdot[VRI_n] + K_{RD}\cdot[PDP_n]}\right)\right) - D_m\cdot[Clk_m] \quad (2.37)$$

$$\frac{\mathrm{d}\,[VRI_c]}{\mathrm{dt}} = T_V\cdot[Vri_m] - N_V\cdot[VRI_c] - D_V\cdot[VRI_c] \quad (2.38)$$

$$\frac{\mathrm{d}\,[VRI_n]}{\mathrm{dt}} = N_V\cdot[VRI_c] - D_V\cdot[VRI_n] \quad (2.39)$$

$$\frac{\mathrm{d}\,[PDP_c]}{\mathrm{dt}} = T_D\cdot[Pdp_m] - N_D\cdot[PDP_c] - D_D\cdot[PDP_c] \quad (2.40)$$

$$\frac{\mathrm{d}\,[PDP_n]}{\mathrm{dt}} = N_D\cdot[PDP_c] - D_D\cdot[PDP_n] \quad (2.41)$$

$$\frac{\mathrm{d}\,[CLK_c]}{\mathrm{dt}} = T_C\cdot[Clk_m] - k_{CC}\cdot[CLK_c] \quad (2.42)$$

$$\frac{\mathrm{d}\,[CLK\cdot CYC_c]}{\mathrm{dt}} = k_{CC}\cdot[CLK_c] + k_{dCP}\cdot[CLK\cdot CYC\cdot P_c] - P_C\cdot[CLK\cdot CYC_c] - N_C\cdot[CLK\cdot CYC_c] \quad (2.43)$$

$$\frac{\mathrm{d}\,[CLK\cdot CYC\cdot P_c]}{\mathrm{dt}} = P_C\cdot[CLK\cdot CYC_c] - k_{dCP}\cdot[CLK\cdot CYC\cdot P_c] + C_{CP}\cdot[CLK\cdot CYC\cdot P_n] - D_{CP}\cdot[CLK\cdot CYC\cdot P_c] \quad (2.44)$$

$$\frac{\mathrm{d}\,[CLK\cdot CYC_n]}{\mathrm{dt}} = N_C\cdot[CLK\cdot CYC_c] - P_C\cdot[CLK\cdot CYC_n] - P_{CC}\cdot[CLK\cdot CYC_n]\cdot([PER\cdot P_n] + [PER\cdot TIM_n]) \quad (2.45)$$

$$\frac{\mathrm{d}\,[CLK\cdot CYC\cdot P_n]}{\mathrm{dt}} = P_C\cdot[CLK\cdot CYC_n]$$

$$+ P_{CC} \cdot [CLK \cdot CYC_n] \cdot ([PER \cdot P_n] + [PER \cdot TIM_n])$$

$$- C_{CP} \cdot [CLK \cdot CYC \cdot P_n] - D_{CP} \cdot [CLK \cdot CYC \cdot P_n] \quad (2.46)$$

$$Light = \begin{cases} 1 & \text{ZT0} < t < \text{ZT12} \\ 0 & \text{ZT12} < t < \text{ZT24}, \quad \text{CT0} < t < \text{CT24} \end{cases} \quad (2.47)$$

This description does not use time delays and explicitly represents the post-translational modifications of PER and CLK. Illumination in light-dark cycles is modeled via Light, defined as a square wave in Equation 2.47. Light acts upon the degradation of cytoplasmic and nuclear TIM. The transcriptional activation kinetics are borrowed from [11], and described in the Supporting Information. FBM is not explicitly represented because the inclusion of this molecule at limiting concentrations did not significantly alter the presented results (unpublished data). As shown in Figure 2.1 (and based on the observations of [41, 97]), the presence of nuclear PER and PER-TIM dimers causes the phosphorylation of CLK. Once phosphorylated, CLK cannot bind to DNA and is either degraded or exported into the cytoplasm where it can be degraded or dephosphorylated. Chemical reaction rate constants are the only adjustable parameters for which a set of biologically meaningful values was found (see Parameter Estimation below). See Table 2.3 for a full list of reaction rate constants for the full circadian model.

2.4.3 Model Solutions

With the exception of the positive feedback model of the isolated *per*/*tim* loop, the mathematical models presented in this paper were solved using the LSODAR integrator as part of the SloppyCell package [29]. Periodic orbits were found through sequential integration cycles until a stable limit cycle was approached. For the continuation analysis of model parameters, AUTO 2000 was used.

Since a small number of molecules may initiate positive feedback, the isolated

per/*tim* loop positive feedback model was solved stochastically using Gillespie's algorithm. The model results are an ensemble of trajectories for a given parameter set (a randomly selected subset which is shown in Figure 2.4B), with the trajectory closest to the experimentally observed mean nuclear onset time used in Figure 2.3A. The standard deviation of nuclear onset time was determined from this ensemble of trajectories.

2.4.4 Mutant Phenotype Characterization

Several *Drosophila* mutant phenotypes were represented by the detailed mathematical model. The parameter changes used to represent the mutants described in the paper are shown in Table 2.4. A typical result is shown for the arrhythmic dbt^{P}/dbt^{ar} mutants in Figure 2.9. The transient trajectory from a point on the wild-type constant darkness limit cycle illustrates the approach to a stable fixed point solution. Similarly, all arrhythmic mutants presented in Table 2.1 were found to approach stable fixed points (unpublished data).

2.4.5 Experimental Data

The points and error bars presented in Figure 2.3 were the result of averaging the five trajectories for cytoplasmic PER-TIM dimers and nuclear PER shown in Figure 2.1B of Meyer et al. [62]. These trajectories were normalized to a minimum of zero and maximum of one prior to aligning the paired PER-TIM and nuclear PER trajectories by minimizing the root mean-squared distance. The mean onset time of the average of the aligned trajectories was then set to 340 min.

The points and error bars in Figure 2.6 were adapted from the publications listed

in Table 2.5. With the exception of *pdp* mRNA and protein, multiple references were available. These datasets were interpolated and averaged to produce the means and standard deviations presented in Figure 2.6. *pdp* mRNA and protein means and error bars were taken directly from [17].

The points and error bars in Figure 2.7 are the average and standard deviation of experimental observations of the period of oscillation in response to changes in *per* dosage [8, 81] and *tim* dosage [5, 72].

2.4.6 Parameter Estimation

A Monte Carlo random walk, guided by importance sampling, adjusted model parameters to optimize a chi-squared value quantifying the consistency of the model results with available experimental observations (discussed in the previous section and presented in Figures 2.3, 2.6, and 2.7). Model parameters were manually adjusted to biologically meaningful values where necessary.

2.4.7 Model Error Bars by an Ensemble Approach

The parameter set found in the previous section is just one of an ensemble of parameter sets consistent with the experimental observation. Using the entire ensemble of parameter sets consistent with experimental observations, uncertainties in the model predictions may be calculated and plotted in Figures 2.6 and 2.7 [30]. These error bars represent the confidence in the ability of the experimentally observed behavior to constrain the model predictions, and thus provide support not only for the particular choice of model parameters but also for the significance of the model results. As a result of the time constraints

of dissertation submission, the incomplete results are not plotted in Figures 2.6 and 2.7, but the methodology is described below.

The estimation of model error bars via an ensemble approach is inspired by the work of Sethna and coworkers [30]. Using a similar approach to the parameter estimation methodology, the ensemble of parameter sets consistent with the experimental observations is generated by a Monte Carlo random walk. To maintain biologically meaningful parameter values, parameter values (v) were loosely constrained to the best fit values (v_0) by adding small prior residuals $\left(\frac{\ln(v/v_0)}{1000}\right)$ to the chi-squared function. The efficiency of the random walk is improved by importance sampling using a Hessian matrix to estimate the parameter sensitivity of the chi-squared function. The Hessian matrix was calculated using either finite differences or the Levenberg-Marquardt algorithm. The duration of the random walk was sufficient to ensure the full exploration of parameter space, indicated by repeated visits to previously explored parameter space. Finally, the model results of interest from a random subset of the ensemble of parameter sets was used to determine the confidence in the model prediction.

2.5 Supporting Information

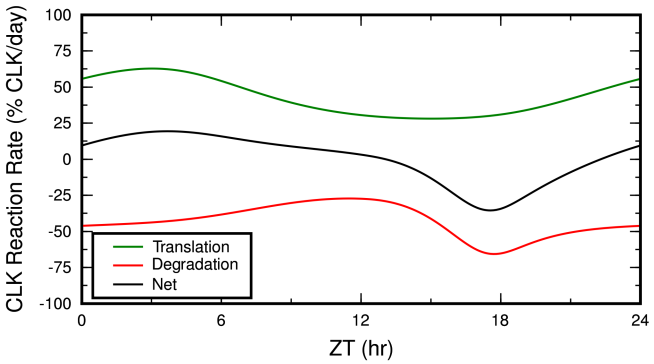

Figure 2.8: The translation and degradation of *clk* protein balance to produce a nearly constant level of total CLK. The reaction rate shown is normalized by the maximum level of total CLK. The degradation term includes both cytoplasmic and nuclear degradation (as defined in Equations 2.44 and 2.46).

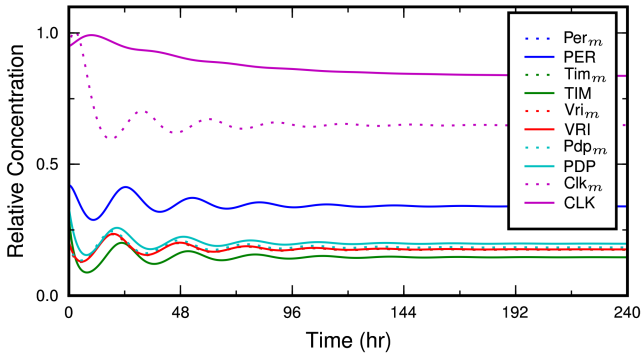

Figure 2.9: The evolution of the dbt^P/dbt^{ar} mutants away from a point on the wild-type constant darkness limit cycle

Table 2.2: Isolated model parameters

Name	Value	Description
k_{MA}	0.2 hr^{-1}	PER-TIM phosphorylation
k_5	0.9091 hr^{-1}	Serial Reaction Coefficient ($N = 5$)
k_{10}	1.818 hr^{-1}	Serial Reaction Coefficient ($N = 10$)
k_{25}	4.545 hr^{-1}	Serial Reaction Coefficient ($N = 25$)
k_{50}	9.091 hr^{-1}	Serial Reaction Coefficient ($N = 50$)
k_{100}	18.18 hr^{-1}	Serial Reaction Coefficient ($N = 100$)
k_{250}	45.45 hr^{-1}	Serial Reaction Coefficient ($N = 250$)
k_{PTc}	0.01 $(nM \cdot hr)^{-1}$	Cytoplasmic PER-TIM association
k_{dPTc}	0.1 hr^{-1}	Cytoplasmic PER-TIM dissociation
k_{PTf}	0.0001 $(nM \cdot hr)^{-1}$	Foci PER-TIM accumulation
P_{PT}	1.6 x 10-7 hr^{-1}	Foci PER-TIM auto-phosphorylation
k_{FB}	0.02 $(nM \cdot hr)^{-1}$	Foci PER-TIM phosphorylation via feedback
k_{SM}	0.25 hr^{-1}	SM_c activation
D_{SM}	1 hr^{-1}	SM_c degradation
N_P	5 hr^{-1}	Phosphorylated PER nuclear transport
C_P	0.01 hr^{-1}	PER cytoplasmic transport
N_T	1 hr^{-1}	TIM nuclear transport
C_T	500 hr^{-1}	TIM cytoplasmic transport
k_{PTn}	1 $(nM \cdot hr)^{-1}$	Nuclear PER-TIM association
k_{dPTn}	0.01 hr^{-1}	Nuclear PER-TIM dissociation

Table 2.3: Full circadian model parameters

Name	Value	Description
I_m	6.027 nM/hr	Maximum transcription rate
D_N	1.363 x 106 nM	Non-specific DNA binding sites
K_{RC}	1.372 x 104	CLK-CYC-DNA relative equilibrium constant
K_{RV}	1.65 x 109	VRI-DNA relative equilibrium constant
K_{RD}	4.007 x 106	PDP-DNA relative equilibrium constant
k_{bP}	54.59 nM/hr	CLK-CYC activation of per
k_{bT}	38.05 nM/hr	CLK-CYC activation of tim
k_{bV}	64.87 nM/hr	CLK-CYC activation of vri
k_{bD}	423.4 nM/hr	CLK-CYC activation of pdp
k_{bC}	29.66 nM/hr	PDP activation of clk
D_m	0.2731 hr^{-1}	mRNA degradation
T_P	0.5686 hr^{-1}	PER translation rate
T_T	1.041 hr^{-1}	TIM translation rate
T_V	4.602 hr^{-1}	VRI translation rate
T_D	0.1649 hr^{-1}	PDP translation rate
T_C	0.06578 hr^{-1}	CLK translation rate
k_{PTc}	33.15 $(nM \cdot hr)^{-1}$	Cytoplasmic PER-TIM association
k_{PTf}	4.092 hr^{-1}	Foci PER-TIM accumulation
P_{PT}	0.0009879 $(nM \cdot hr)^{-1}$	Foci PER-TIM auto-phosphorylation
k_{PTn}	6.668 $(nM \cdot hr)^{-1}$	Nuclear PER-TIM association
k_{dPTn}	1.483 hr^{-1}	Nuclear PER-TIM dissociation
k_{SM}	1.484 hr^{-1}	SM_c activation
k_{FB}	1.473 $(nM \cdot hr)^{-1}$	Foci PER-TIM phosphorylation via feedback

Table 2.3 (Continued)

Name	Value	Description
k_{CC}	2.497 hr^{-1}	CLK association to CYC
k_{dCP}	0.4666 hr^{-1}	CLK dephosphorylation
N_P	26.08 hr^{-1}	Phosphorylated PER nuclear transport
N_T	8.418 hr^{-1}	TIM nuclear transport
N_V	3.829 hr^{-1}	VRI nuclear transport
N_D	0.4847 hr^{-1}	PDP nuclear transport
N_C	0.1257 hr^{-1}	CLK nuclear transport
C_{CP}	3.058 hr^{-1}	Phosphorylated CLK cytoplasmic transport
P_{CC}	23.97 $(nM \cdot hr)^{-1}$	PER-mediated CLK phosphorylation
P_C	0.01005 hr^{-1}	non-PER-mediated CLK phosphorylation
D_P	0.5892 hr^{-1}	PER degradation
D_T	1.252 hr^{-1}	TIM degradation
D_V	2.481 hr^{-1}	VRI degradation
D_D	0.4984 hr^{-1}	PDP degradation
D_{CP}	0.0975 hr^{-1}	Phorphorylated CLK degradation
D_{SM}	0.7384 hr^{-1}	SM_c degradation
D_L	0.2276 hr^{-1}	Light-activated degradation

Table 2.4: Parameter changes to describe mutant phenotypes

Mutant	Parameter	Reference
per^{01}	$T_P = 0$ hr^{-1}	[8]
tim^{01}	$T_T = 0$ hr^{-1}	[64]
dbt^{P}, dbt^{ar}	$D_P = 0.1736$ hr^{-1}	[68, 73]
clk^{Jrk}	$T_C = 0$ hr^{-1}	[3]
cyc^{0}	$k_{CC} = 0$ hr^{-1}	[75]
pdp^{P205}	$T_D = 0$ hr^{-1}	[17]

Table 2.5: Experimental mRNA and protein measurement sources

Molecule	Reference(s)
per mRNA	[6, 7, 32, 40, 84, 98]
PER	[6, 47, 57, 98]
tim mRNA	[7, 9, 40, 77, 84]
TIM	[6, 47, 57]
clk mRNA	[7, 17, 25, 40]
vri mRNA	[9, 17]
VRI	[17, 25]
pdp mRNA	[17]
PDP	[17]

2.5.1 Derivation of Transcription Activation Terms

The activation of *per*, *tim*, *vri*, and *pdp* transcription are derived as described in the Supporting Text of Bolouri and Davidson [11]. The activation of *clk* transcription by *pdp* with repression by *vri* is derived below. Similar to Bolouri and Davidson, we assume: 1) equilibrium in factor binding to both specific and non-specific sites, 2) essentially all factor is bound to either specific or non-specific sites (i.e. negligible free factor), and 3) total factor concentration is much greater than specific site-bound factor (or similarly that non-specific site-bound factor greatly outnumbers specific site-bound factor). Finally, we assume VRI competitively inhibits PDP binding.

$$Y = \frac{[PDP_n \cdot D_S]}{[D_S] + [VRI_n \cdot D_S] + [PDP_n \cdot D_S]} \tag{2.48}$$

$$K_{equilibrium\_S,VRI} = \frac{[VRI_n \cdot D_S]}{[D_S] \cdot [VRI_n]} \tag{2.49}$$

$$K_{equilibrium\_N,VRI} = \frac{[VRI_n \cdot D_N]}{[D_N] \cdot [VRI_n]} \tag{2.50}$$

$$K_{RV} = \frac{K_{equilibrium\_S,VRI}}{K_{equilibrium\_N,VRI}} = \frac{[VRI_n \cdot D_S]}{[D_S]} \cdot \frac{[D_N]}{[VRI_n \cdot D_N]} \tag{2.51}$$

Similarly for PDP binding:

$$K_{RD} = \frac{K_{equilibrium\_S,PDP}}{K_{equilibrium\_N,PDP}} = \frac{[PDP_n \cdot D_S]}{[D_S]} \cdot \frac{[D_N]}{[PDP_n \cdot D_N]} \tag{2.52}$$

Applying assumptions 2 and 3 yields:

$$[VRI_n] = \left[VRI_{n,free}\right] + [VRI_n \cdot D_S] + [VRI_n \cdot D_N] = [VRI_n \cdot D_N] \tag{2.53}$$

$$[PDP_n] = \left[PDP_{n,free}\right] + [PDP_n \cdot D_S] + [PDP_n \cdot D_N] = [PDP_n \cdot D_N] \tag{2.54}$$

$$Y = \frac{K_{RD} \cdot [PDP_n]}{[D_N] + K_{RV} \cdot [VRI_n] + K_{RD} \cdot [PDP_n]} \tag{2.55}$$

The final kinetic term is derived as described in the Supplemental Text of Bolouri and Davidson [11].

2.6 Acknowledgments

We would like to thank John Ewer and John Guckenheimer for their advice in constructing the mathematical model and Jim Sethna for his helpful discussions and his assistance developing an algorithm to find periodic limit cycles. We also thank Venus So and Michael Rosbash for raw data, and Justin Blau for helpful discussions. JGO acknowledges support from the Ministerio de Educación y Ciencia (Spain) and from the Generalitat de Catalunya.

CHAPTER 3

THE COMMUNICATION UNDERLYING SEASONAL ADAPTATION OF *DROSOPHILA* CIRCADIAN RHYTHMICITY

In this study we present an investigation of the circadian pacemaker coupling underlying the adaptation of *Drosophila* locomotor activity to seasonal changes in day length. Building on our prior detailed model of *Drosophila* circadian rhythmicity, the possible inter-pacemaker signal generation and signaling targets are systematic searched for combinations consistent with experimentally observed behavior. The results of this search are compared against the experimentally observed codominance of the morning and evening oscillators thought to give rise to *Drosophila* seasonal adaptation. Further experimentation is suggested to better define these interactions and improve the understanding and control of circadian rhythms.

3.1 Introduction

Circadian rhythmicity is a fundamentally interesting biological phenomenon which has important impacts on human health. Two such circadian health impacts are jet lag, thought to arise from poorly synchronized pacemakers [95], and chemotherapy, which may exploit the loss of normal circadian cell proliferation control [22]. These examples illustrate the importance of understanding not only the circadian pacemakers, but the resulting communication from (and communication between different) circadian pacemakers. A widely studied circadian output is the crepuscular locomotor activity of the fly *Drosophila*, which adapts to seasonal changes in day length. While long suspected, recent experimental evidence now supports the codominance of the morning and evening pacemakers giving rise to seasonal adaptation [89]. Thus, the seasonal adaptation of the

fruit fly provides a model system for studying circadian communication where insights may also improve our understanding of the similarly structured mammalian circadian clock.

Previous mathematical models of circadian rhythms have primarily focused on the individual neuron and the properties of its oscillation [44, 50, 76, 83, 91, 92]. Relatively few models have been extended to the investigation of coupled *Drosophila* oscillators. One example is the work of Ueda and coworkers investigating the possible targets of light-pulse and dark-pulse phase shifts [93]. While providing significant insight into possible mechanisms, the author's clock architecture did not capture the heterogeneity of pacemaker types or describe circadian outputs (e.g. locomotor activity). Here we present a computational study of the possible interactions underlying the seasonal adaptation of *Drosophila* locomotor activity. A systemic search is used to find the clock communication mechanism most consistent with the observed codominant morning and evening pacemakers. The results of this search illustrate possible signaling mechanisms, are compared against recent experimental observations, and used to suggest possible future experimentation.

3.2 Methods

The mathematical model developed below is available online [45].

3.2.1 Mathematical Model of an Individual Neuron

The circadian network used to describe the behavior of individual neurons was a modification of our previously described model [44]. Unless otherwise noted, the following

description uses the kinetics and parameters described therein. New parameter values are defined in the Supplementary Material. The *Drosophila* circadian clock is composed of two interlocking feedback loops. The first loop is composed of the negative feedback of *period* (*per*) and *timeless* (*tim*), which down-regulate their own expression by inhibiting the CLOCK-CYCLE (CLK-CYC) transcription factor. PER dimerizes with TIM before localizing to the nucleus via an uncharacterized mechanism. The SHAGGY (SGG) kinase has been shown to phosphorylate TIM, also affecting the nuclear localization of PER and TIM. Circadian rhythms are entrained by light through an increased degradation of TIM protein. Light was represented as a square wave with 12 hours of light and 12 hours of dark. Winter and summer days had 10 and 14 hours of light, respectively, with the balance of 24 hours dark. In the second feedback loop, the expression of *clk* is regulated by *vrille* (*vri*) and *PAR domain protein 1 isoform* ϵ (*pdp*). Both *vri* and *pdp* expression are activated by CLK-CYC. VRI represses the expression of *clk*, creating a negative feedback loop, whereas PDP creates a positive feedback loop through activating *clk* expression.

The kinetics of *clockwork orange* (*cwo*) were added to capture current experimental observations [39, 51, 59]. The transcription of *cwo* mRNA (Cwo_m), the translation of cytoplasmic protein (CWO_c), and the transport of active nuclear protein (CWO_n) were described by the following set of equations.

$$A_{Ebox} = \frac{K_{RC} \cdot [CLK \cdot CYC_n]}{DN + K_{RW} \cdot [CWO_n] + K_{RC} \cdot [CLK \cdot CYC_n]} \tag{3.1}$$

$$\frac{\mathrm{d}\,[Cwo_m]}{\mathrm{dt}} = I_m \cdot \left(1.0 - \exp\left(-\left(\frac{k_{bW}}{I_m} A_{Ebox}\right)\right)\right) - D_m \cdot [Cwo_m] \tag{3.2}$$

$$\frac{\mathrm{d}\,[CWO_c]}{\mathrm{dt}} = T_W \cdot [Cwo_m] - N_W \cdot [CWO_c] - D_W \cdot [CWO_c] \tag{3.3}$$

$$\frac{\mathrm{d}\,[CWO_n]}{\mathrm{dt}} = N_W \cdot [CWO_c] - D_W \cdot [CWO_n] \tag{3.4}$$

Similarly, the transcription of other E-box regulated genes (*per*, *tim*, *vri*, and *pdp*) included the $K_{RW} \cdot [CWO_n]$ term shown in Equation 3.1 to reflect CWO repression

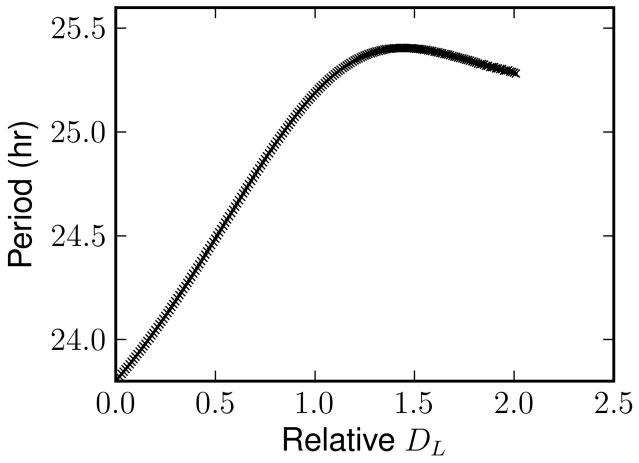

Figure 3.1: Light sensitivity of the circadian clock and its effect on the period of oscillation in constant light. Low light sensitivity shows a period-elongating response to increased light consistent with evening oscillators whereas high light sensitivity shows a period-shortening response consistent with morning oscillators.

[39, 51, 59].

Next, the above description was refined by the unique differences of the morning and evening oscillators of the classic Pittendrigh and Dann model [28, 67]. With the exception of the dynamics of intercellular signaling and light sensitivity, both the morning and evening oscillators were assumed to share the above circadian clock kinetics and parameters. The primary difference between the two oscillator types was assumed to be the greater light sensitivity of the morning oscillator compared to the evening oscillator [70]. Additionally, the morning oscillator is observed to have a period-shortening response to light whereas the evening oscillator elongates its period [70], which is consistent with the light sensitivity results from our previous model [44] (Figure 3.1). To represent these observations, the light sensitivity parameter (formerly D_L) was made oscillator type specific, with light sensitivity in the morning and evening oscillators described by D_{LM} and D_{LE}, respectively.

3.2.2 Coupling of Evening and Morning Oscillators

Pacemaker communication is required for behavioral rhythms in constant darkness [52], with the neuropeptide *pigment dispersing factor* (*pdf*) the best characterized signal in the *Drosophila* clock [69]. *pdf* is expressed in morning oscillators, activated by a PER- / TIM-dependent mechanism [66], and received by the PDF-receptor present on evening oscillators [37, 46, 61]. Based on these observations, PDF was assumed to be the neural signal communicating the phase from the morning oscillator to the evening oscillator. From a comparison of the observed phase of PDF expression [66], the positive feedback molecule SM_c was used to activate PDF.

$$\frac{\mathrm{d}\,[PDF_c]}{\mathrm{dt}} = k_{PDF} \cdot [SM_c] - E_{PDF} \cdot [PDF_c] - D_{PDF} \cdot [PDF_c] \tag{3.5}$$

$$\frac{\mathrm{d}\,[PDF_e]}{\mathrm{dt}} = E_{PDF} \cdot [PDF_c] - D_{PDF} \cdot [PDF_e] \tag{3.6}$$

After activation, PDF is exported extracellularly from the cytoplasm and degraded as described by the E_{PDF} and D_{PDF} reaction rate constants, respectively. Because the ultimate circadian input of PDF on the evening oscillators is unknown, possible reactions were searched using the methodology described below.

Under constant light conditions, the evening oscillator may be shown to be the primary pacemaker [63] and may control the phase of the morning oscillator during long summer days [89]. Unfortunately, little evidence is available to define the underlying neuropeptide or electrical signal responsible for this communication. The reciprocal communication of phase from evening to morning cells was thus named FDP (the reverse of PDF) and assumed to be produced in evening oscillators and act on morning oscillators. If we also assume FDP is responsible for the evening locomotor activity peak, the expression of FDP is consistent with being, or being regulated by, an E-box

regulated gene.

$$\frac{\mathrm{d}\,[FDP_c]}{\mathrm{dt}} = k_{FDP} \cdot I_m \cdot \left(1.0 - \exp\left(-\left(\frac{k_{bP}}{I_m} A_{Ebox}\right)\right)\right) - E_{FDP} \cdot [FDP_c] - D_{FDP} \cdot [FDP_c] \quad (3.7)$$

$$\frac{\mathrm{d}\,[FDP_e]}{\mathrm{dt}} = E_{FDP} \cdot [FDP_c] - D_{FDP} \cdot [FDP_e] \quad (3.8)$$

Similar to PDF, the target of FDP signaling is unknown and thus possible reactions were searched using the methodology described below.

Extracellular PDF and FDP were assumed to create the observed morning and evening locomotor activity peaks, respectively. The cooperative nature of PDF and FDP cell receptor binding and signaling was represented by an exponent of 4 in the expression for the locomotor activity (LMA):

$$LMA = [PDF_e]^4 + k_{LMA} \cdot [FDP_e]^4 \quad (3.9)$$

where k_{LMA} is used to balance the relative contributions of $[PDF_e]$ and $[FDP_e]$.

After defining the unknown reactions for PDF and FDP signaling, a system of differential equations was compiled to describe the circadian behavior of the morning and evening oscillators. The system was solved deterministically using the DDASKR integrator of the SloppyCell package [29], implicitly assuming moderate or large molecule concentrations. As a result of using a deterministic solution, only one morning and one evening oscillator were necessary. Periodic orbits were found through sequential integration until a stable periodic solution was approached.

3.2.3 *shaggy* Over-Expression

In the experimental observations of the seasonal adaptation of locomotor activity presented below, Stoleru and coworkers [89] used *sgg* over-expression to selectively ac-

celerate morning or evening oscillators. The *sgg* kinase has been observed to phosphorylate TIM, with *sgg* over-expression producing a shortened period [58]. Beyond the advanced nuclear localization of PER-TIM with *sgg* over-expression, little is known how this kinase accelerates rhythmicity. Thus, a search of possible reaction mechanisms was performed as described below.

3.2.4 Parameter and Reaction Estimation

All model parameters were estimated using a Monte Carlo random walk. Importance sampling was used when the number of parameters searched was less than ten, otherwise uniform sampling was used. As a result of the artificial time constraints of dissertation submission, confidence intervals have not been assigned to the model predictions, but may be calculated using the methodology described in Section 2.4.7.

The unknown reactions for the inter-pacemaker signaling and *sgg* over-expression were found using an automated search inspired by [12], although reaction kinetic laws were not directly altered but instead selected from a list of possible mechanisms. The kinetic laws were not directly altered here because the resulting kinetics frequently produced unstable oscillations. The possible reaction mechanisms were searched using a genetic algorithm to find the combination of reactions most consistent with the experimental observations described below. The mechanisms of three reactions were searched, including the target of PDF and FDP signaling and the effect of *sgg* over-expression. These reactions were assumed to catalytically modify existing model kinetics by regulating either specific reactions or classes of reactions via a model parameter. For example, selecting the PDF target as regulating *per* translation would replace the $T_P \cdot [Per_m]$ kinetic law with $(1 + v_{PDF} \cdot [PDF_e]) \cdot (T_P \cdot [Per_m])$. The genetic algorithm used a pop-

ulation size of 40 individuals per generation with 8 parents chosen from tournaments of 10 individuals for the next generation. Each new child was equally probable to be a mutant of one parent (randomly changing a single reaction) or the progeny of two randomly selected parents (resulting in a random combination of the parent reactions). After selecting the reactions for each child, a stable solution was found (if none was found, the child was discarded) and the parameters for PDF and FDP signaling and *sgg* over-expression were re-estimated.

The interaction of morning and evening oscillators was partitioned for searching reaction and parameter space. This partitioning was useful to delineate the cause and effect of the circadian communication mechanisms present in a circular network of PDF and FDP. To partition the morning oscillator, the experimentally observed evening locomotor activity peak [89] was assumed to represent extracellular FDP with the expression of extracellular PDF compared to the morning locomotor activity peak. Similarly, the evening oscillators were partitioned using morning locomotor activity peak [89] to represent extracellular PDF with FDP compared to the evening locomotor activity peak. After searching reaction space using partitioned morning and evening oscillators, the results were manually combined and the parameters re-estimated.

3.2.5 Experimental Data

A least-squares cost function quantified the consistency of the model results with the wild-type mRNA and protein observations previously described [44], wild-type *cwo* mRNA observations [39, 51, 59], and experimentally observed locomotor activity. The locomotor activity profiles used include the wild-type and *sgg* over-expressing mutants under winter and summer conditions observed by Stoleru and coworkers [89]. The

relative errors of mRNA / protein data and locomotor activity data were balanced to maintain approximately equal contributions to the final cost.

3.3 Results

3.3.1 Coupling Reactions

The reaction mechanism search results are summarized in Table 3.1 with the reactions selected for the final results presented below. The PDF signaling circadian input was found to be the turnover of PDP or nuclear PER. Because PER activity is more central to clock oscillations, the degradation of nuclear PER was selected. Thus, the $D_P \cdot [PER \cdot P_n]$ term was replaced by $(1 + v_{PDF} \cdot [PDF_e]) \cdot (D_P \cdot [PER \cdot P_n])$ in the evening oscillators. The FDP signal was also found to target nuclear PER stability or PDF activation. The activation of PDF was selected to increase the dominance of the evening oscillator over the morning activity peak, replacing the $k_{PDF} \cdot [SM_c]$ term of Equation 3.5 with $(1 + v_{FDP} \cdot [FDP_e]) \cdot (k_{PDF} \cdot [SM_c])$ in morning oscillators. Finally, the over-expression of *sgg* targeted several phase-advancing reactions: advancing PER-TIM nuclear transport via early cytoplasmic dimerization, increasing CLK levels (by controlling *clk* transcription), and early CLK nuclear transport. Because of prior experimental observations of SGG affecting TIM phosphorylation [58], the cytoplasmic dimerization of PER-TIM was selected. As a result, the $k_{PTc} \cdot [PER_c] \cdot [TIM_c]$ term was replaced by $k_{SGG} \cdot (k_{PTc} \cdot [PER_c] \cdot [TIM_c])$ in *sgg* over-expressing pacemakers.

Table 3.1: Results of the circadian communication mechanism search

Reaction	**Effect**
PDF Signal	**Degrade nuclear PER**
	Unbind nuclear PER-TIM
	Degrade PDP
FDP Signal	**Increase PDF activation**
	Increase light sensitivity of nuclear PER-TIM
SGG Over-Expression	**Increase cytoplasmic PER-TIM binding**
	Increase CLK-CYC nuclear transport
	Decrease cytoplasmic PDP degradation
	Decrease *vri* transcription

Notes: Bold reactions were selected for the final results.

3.3.2 Coupled Oscillations

The above signaling reactions were implemented to couple morning and evening oscillators and describe *sgg* over-expression. The parameters defining these newly added kinetics are presented in the Supplementary Material. The stable 12 hr light:12 hr dark oscillations of *per*, *tim*, *clk*, *vri*, *pdp*, and *cwo* mRNA and protein are shown in Figure 3.2. The results show the phase advance of the evening oscillator mRNA and protein (dashed) with respect to the morning oscillator mRNA and protein (dotted). Additionally, the results are consistent with experimentally observed modest oscillations in total CLK [41, 97] and *cwo* mRNA [39, 51, 59]. Under constant darkness conditions, a stable oscillation was found with a period of 23.8 hours.

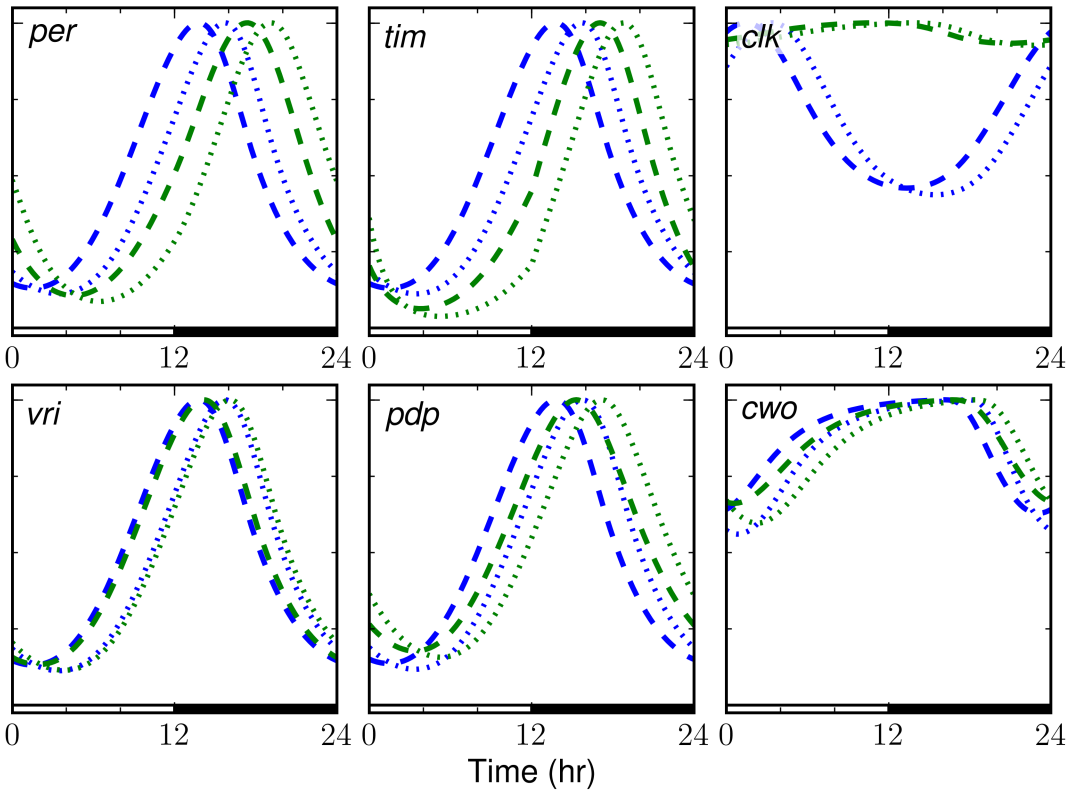

Figure 3.2: Trajectories of mRNA (blue) and protein (green) for the central clock proteins in the morning (dotted) or evening (dashed) oscillators. The open and closed bar represent light and dark conditions, respectively.

3.3.3 Seasonal Adaptation

The predicted seasonal adaptation of locomotor activity is compared to the experimentally observed profiles [89] in Figure 3.3 for specific over-expression of *sgg* in morning oscillators, evening oscillators, or neither. The model results show similar shifts of about 2.5 hours in the morning and evening locomotor activity peaks between winter and summer conditions while maintaining an approximately 13 hour delay between the morning and evening activity peaks. The over-expression of *sgg* shows the ability to specifically shift the corresponding activity peak in both the summer and winter conditions.

The model results are not able to capture all characteristics of the experimentally

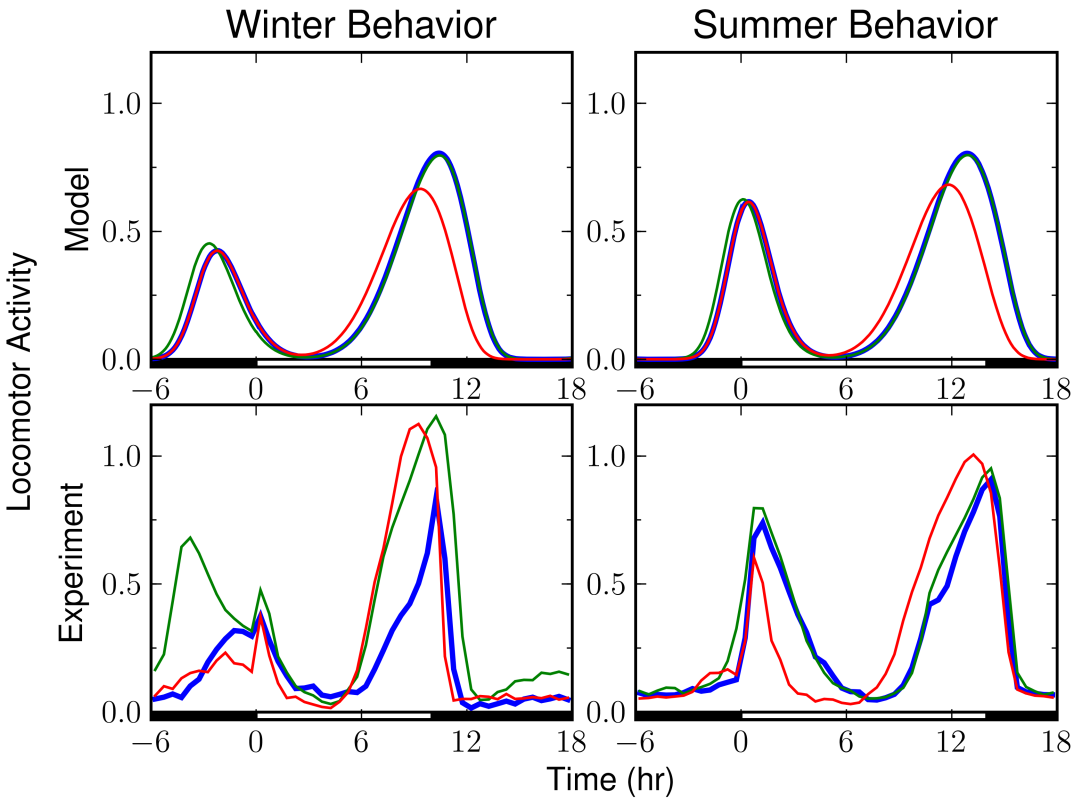

Figure 3.3: Seasonal adaptation of mathematical model (top panels) compared to experimental observations (bottom panels) [89]. Wild-type (blue), morning oscillator *sgg+* (green), and evening oscillator *sgg+* (red) genetic conditions are shown. Winter conditions (10 hours of light) and summer conditions (14 hours of light) are indicated by the open and closed bars.

observed seasonal adaptation [89]. Primarily, neither the morning nor the evening oscillator appears to be dominant over the other in the winter or summer conditions. While the circadian communication leads to differences in gene expression, the effect of *sgg* over-expression has little effect on the complementary locomotor activity peak.

3.4 Discussion

The results of a search for the circadian communication responsible for the seasonal adaptation of *Drosophila* locomotor activity illustrates a novel signaling network connecting morning and evening oscillators. The activation of the well characterized PDF neuropeptide is consistent with the phase of the feedback responsible for PER switch-like dissociation from TIM [44] in morning oscillators. The activation of the opposing hypothetical FDP signal, responsible for communicating the evening oscillator phase and locomotor activity peak, is found to be consistent with E-box regulated expression in evening oscillators. The resulting targets of the PDF and FDP signals are found to be nuclear PER stability and PDF activation, respectively, in the opposing oscillator types. While other similar reactions certainly make this coupling mechanism one of several possibilities, the results highlight a biologically reasonable coupling mechanism involving well-established circadian regulators including E-box activation and nuclear PER stability.

The seasonal adaptation resulting from circadian communication between morning and evening oscillators is consistent with the ability of each oscillator type to affect its own locomotor activity peak. Additionally, the wild-type phase of the morning and evening locomotor activity peaks show good agreement with experimental observations [89]. Unfortunately, the model results do not seem capable of capturing the codominance of the morning and evening oscillators in different seasons as shown by the ability of each oscillator type to drive the complementary activity peak. Our inability to establish this codominance may represent an incomplete search of reaction space, possibly a result of partitioning morning and evening oscillators, or may represent a gap in understanding circadian regulation.

Several experiments could confirm or invalidate the signaling mechanisms proposed here. Studying PER nuclear staining in the evening oscillators should show lowered staining with *pdf* over-expression in the morning oscillators. Additionally we find SGG advances PER-TIM cytoplasmic association, which could be tested by a FRET experiment (similar to [62]) used to elicit whether *sgg* specifically affects the binding and / or nuclear transport PER-TIM. Finally, a search of E-box regulated genes which share homology to known neuropeptides may yield the identity of FDP.

3.5 Conclusions

Many interesting circadian-associated phenomenon are a result of pacemaker coupling. Here we present a systematic search of parameter and reaction space for the coupling present in the *Drosophila* circadian clock. Our results show the putative evening signal may be a product of E-box transcription and affect PDF activation. Conversely, PDF is produced in a PER-dependent manner and our results show it affects nuclear PER stability. This coupling mechanism is consistent with many of the characteristics of seasonal adaptation of locomotor activity to day length but does not capture the codominance of different pacemaker types. Thus, the mechanism presented here is used to suggest some future experimentation to further elicit the underlying circadian communication.

3.6 Supplementary Material

Table 3.2: Model Parameter Values

Parameter	Value	Description
k_{bW}	4.84×10^4 nM/hr	CLK-CYC activation of *cwo*
K_{RW}	675	CWO-DNA relative equilibrium constant
T_W	0.967 hr^{-1}	CWO translation rate
N_W	0.349 hr^{-1}	CWO nuclear transport
D_W	0.645 hr^{-1}	CWO degradation
v_{PDF}	0.0801 hr^{-1}	PDF signaling
k_{PDF}	85.1 hr^{-1}	PDF cytoplasmic production
E_{PDF}	124 hr^{-1}	PDF extracellular secretion
D_{PDF}	71.5 hr^{-1}	PDF degradation
v_{FDP}	0.0278 hr^{-1}	FDP signaling
k_{FDP}	829 hr^{-1}	FDP cytoplasmic production
E_{FDP}	46.9 hr^{-1}	FDP extracellular secretion
D_{FDP}	76.9 hr^{-1}	FDP degradation
k_{SGG}	1.19 hr^{-1}	SGG over-expression
k_{LMA}	1.05	LMA scaling coefficient for FDP_e
D_{LM}	0.38 hr^{-1}	Light-activated degradation (Morning)
D_{LE}	0.139 hr^{-1}	Light-activated degradation (Evening)

CHAPTER 4

CELL-WIDE MODEL OF HETEROLOGOUS GENE EXPRESSION IN *ESCHERICHIA COLI*

Genome-wide measurements provide powerful insights for understanding cellular phenomena. Unfortunately, not all -omic measurements are created equal, especially evident in the cost and incomplete genome coverage of proteomic versus transcriptomic assays. As a result, transcript measurements are commonly assumed to reflect changes in the proteome despite current observations suggesting a complex and nonlinear relationship. Motivated by our group's effort measuring cellular responses in both the transcriptome and proteome, here we present a cell-wide mathematical model of protein synthesis capable of predicting the proteome from transcript measurements. The model results highlight the complex and nonlinear relationship of changes in the transcriptome and proteome. The predicted proteome is consistent with the experimentally observed proteome response to the over-expression of *rhsA* in *Escherichia coli*. Further investigation of genes which are not well-described by the model results illustrates possible additional post-transcriptional regulation, yielding insight into *rhsA* action. Thus the model provides a useful tool for proteomic predictions while supporting the need to either directly measure protein expression or to more carefully address any assumed transcriptome-proteome relationship.

4.1 Introduction

Protein synthesis is a fundamental cellular process, producing much of the cell's functional content from enzymes to regulators. While measuring protein expression is important to determine the cellular state, the current proteome measurement techniques are

expensive in time and resources and lack the genome-wide coverage of transcriptomic measurements. As a result transcript changes are commonly assumed either implicitly or explicitly to reflect changes in the proteome, an assumption which appears to contradict the complex and nonlinear changes in mRNA and protein observed across several organisms including *Escherichia coli* [48, 49], yeast [31, 38, 71], and the human liver [4].

In an attempt to describe the complex process of protein synthesis, numerous mathematical models have been developed [24, 80, 100]. These models incorporate diverse levels of complexity: from a detailed translation elongation description [100] to coarse-grained translation elongation [24, 80] to other translation regulators such as ribosome recycling [24]. These models are primarily applied to the prediction of the specific protein production rate of an individual gene, with relatively few works extending the model description to cell-wide protein synthesis. One cell-wide study of protein synthesis is that of Mehra and Hatzimanikatis [60], although their results were based upon a hypothetical *E. coli* transcriptome using uniform translation initiation and elongation rates. Additionally, the deterministic solution used may have numerical issues [35, 79] which could impact the accuracy of their calculated ribosome distributions for longer genes with higher ribosome density.

Here we present a mathematical model of cell-wide protein synthesis capturing the competition for cellular resources (i.e. ribosomes) by the *E. coli* transcriptome experimentally observed prior to, and following, the over-expression of *rhsA*. This model uses inhomogenous translation initiation and elongation kinetics for the approximately 4000 genes present in the transcriptome. The model results are compared to the experimentally observed proteome and highlight the significant time delay in the proteomic response to transcriptional change. Additionally, genes which are not well described

by the model results do not appear randomly distributed, but represent relatively few metabolic pathways, consistent with the presence of additional post-transcriptional regulation not present in the model. Thus the model results provide insight into the effects of *rhsA* over-expression while illustrating a possible path for future model development.

4.2 Methods

4.2.1 Experimental Observations

The development of our mathematical model is motivated by ribosome, transcriptome, and proteome measurements of *E. coli* over-expressing the ORF-ex and dsORF-a1 fragments of *rhsA* under the control of an isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible promotor. mRNA and protein concentrations were measured at four time points: a control immediately prior to induction (Control) and at 3.5 hours afterwards with no induction (No), low induction with 0.1 mM IPTG (Low), and high induction at 1.0 mM IPTG (High) [1, 2], as shown in Figure 4.1. The term "concentration" used in this work reflects the number of molecules per cell, unless otherwise noted, and is interchangeable with "expression" and "expression level" which are also commonly used in the literature. Because the transcriptomic and proteomic protocols used provide relative measurements, the experimentally observed microarray signal intensities or mass spectrum peak areas must be scaled to actual concentrations, as described below. Additionally, ribosome elution profiles were collected at the No, Low, and High data points [2]. From the peak areas of these elution profiles the ratio of active (70S and polysome fractions) to inactive 100S ribosomes (inactive 70S dimers) was found.

4.2.2 Growth Rate

Almost immediately following induction of *rhsA*, the *E. coli* cell culture begins to enter stationary phase. The slowing growth rate may reflect changes in cellular physiology affecting the concentration and rates of many model substrates and reactions. To quantify these changes, a logistic equation is used to fit the growth rate (μ), in doublings per hour:

$$\frac{dX_c}{dt} = k \cdot X_c(t) \cdot \left(1 - \frac{X_c(t)}{X_\infty}\right) \tag{4.1}$$

$$\begin{aligned} \mu_c(t) &= \frac{1}{\ln 2} \frac{1}{X_c(t)} \frac{dX_c}{dt} \\ &= \frac{k}{\ln 2} \cdot \left(1 - \frac{\frac{X_0}{X_\infty} \cdot \exp(k \cdot t)}{1 - \frac{X_0}{X_\infty}(1 - \exp(k \cdot t))}\right) \end{aligned} \tag{4.2}$$

where X_0 is the initial biomass concentration, X_∞ is the carrying capacity, k the maximum growth rate in hr^{-1}, and t is the hours since the induction of *rhsA* with IPTG. The subscript c denotes the induction level to which k, X_0, and X_∞ were fit (values shown in the legend of Figure 4.1). The initial growth rate for all culture conditions was constrained to be equal to establish a common initial state. Because biomass concentration is not directly used in the following model development, the initial biomass concentration was not constrained to more accurately capture the growth dynamics.

4.2.3 Translation Substrates

mRNA expression data were collected as described in [2]. Briefly, the mean signal intensity of each gene was found from four normalized replicates (two biological, two technical) at each experimental data point. After excluding open reading frames with in-frame stop codons or frameshifting, a set of 4117 open reading frames represented the *in silico* transcriptome. To scale the signal intensities to actual concentrations, the total

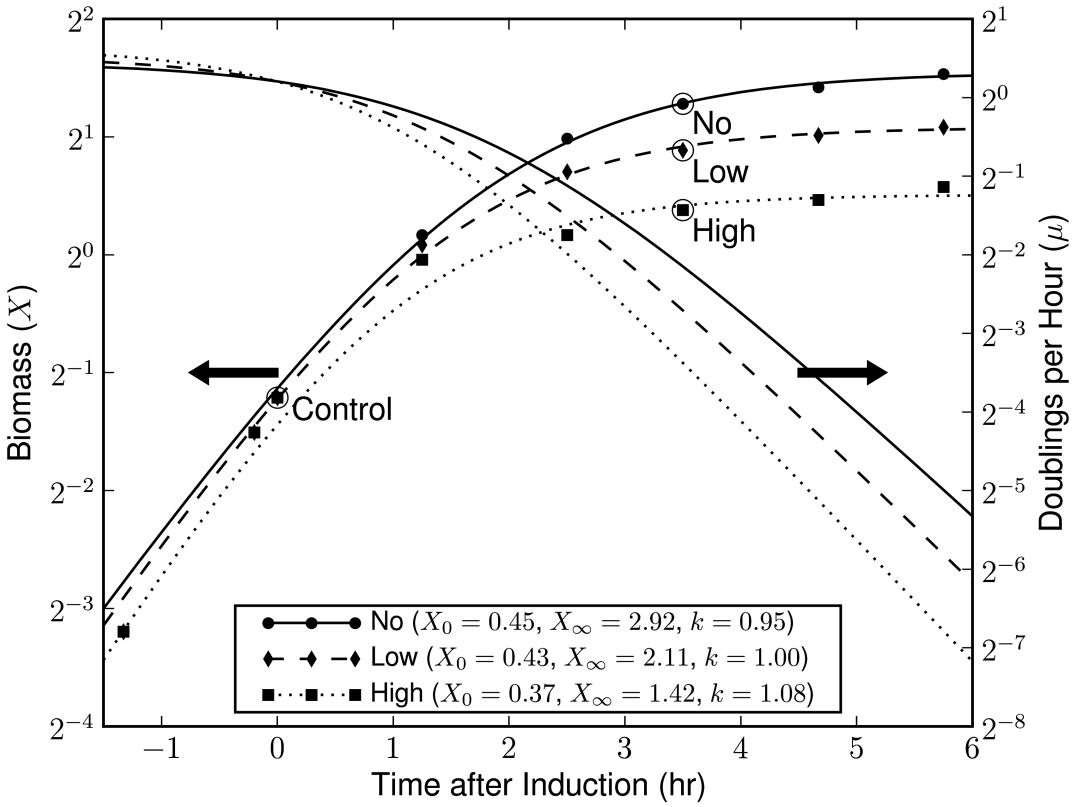

Figure 4.1: Batch culture growth, measured as optical density at 600nm (points), fit by a logistic growth function for each of the three levels of *rhsA* induction. The parameters, defined in the text, are shown in the legend for each culture condition. The four data collection points are labeled Control, No, Low, and High.

mRNA nucleotide concentration at each experimental condition ($N_{NT,c}$) was found from the experimentally observed nucleotide synthesis rate (r_m), a function of the growth rate, subject to a constant degradation rate ($k_{d,m} = 1 min^{-1}$) [13]:

$$\frac{dN_{NT,c}}{dt} = r_m(t) - k_{d,m} \cdot N_{NT,c}(t) = 9.25 \times 10^5 \cdot \mu_c(t) - k_{d,m} \cdot N_{NT,c}(t) \tag{4.3}$$

As shown in Figure 4.2, the nucleotide synthesis rate (r_m) can be linearly fit to the data given at different growth rates in Table 3 of Bremer and Dennis [13]. The individual mRNA concentrations at each experimental data point ($M_{i,c}$) were calculated from the total nucleotide concentration, the gene lengths (L_i), and microarray signal intensities

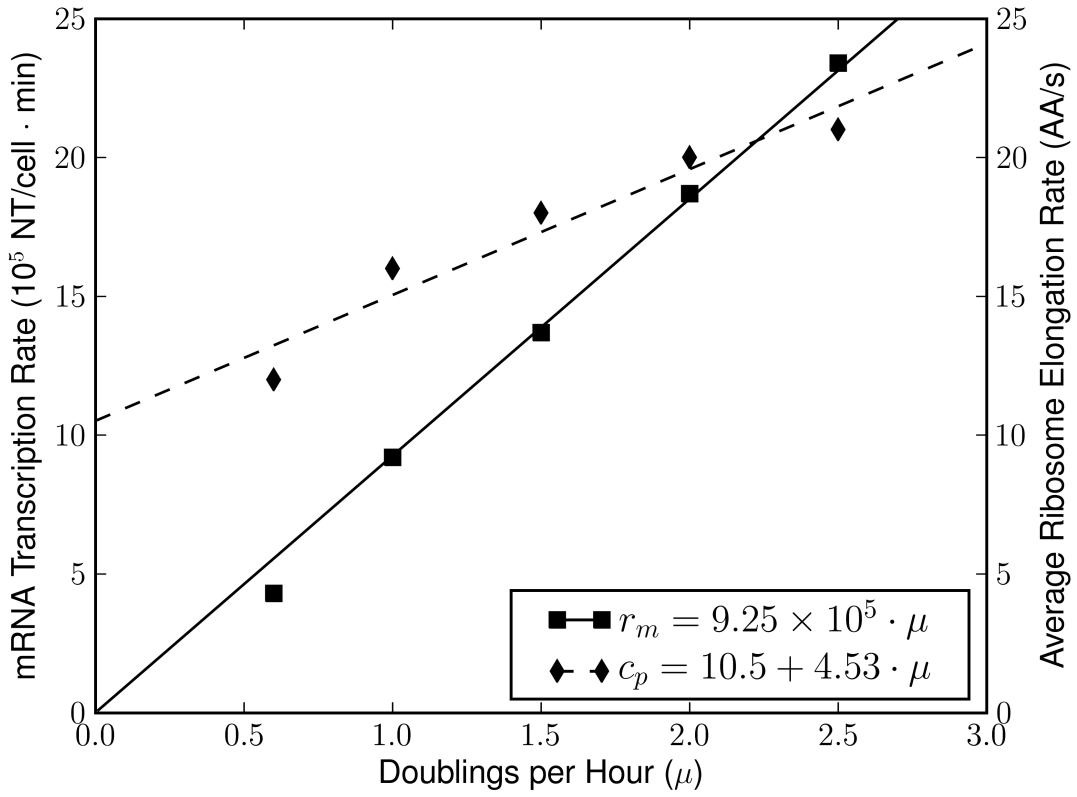

Figure 4.2: mRNA synthesis rate (r_m) versus growth rate from Table 3 of Bremer and Dennis [13] (squares) was fit linearly with a zero intercept to fix the overall mRNA concentration. Similarly, the average ribosome elongation rate (c_p) from the same source [13] (diamonds) was fit linearly with a non-zero intercept and used to estimate changes in translation elongation. The non-zero intercept of this specific (per ribosome) elongation rate reflects the productivity maintained at lower growth rates by the dropping total ribosome concentration.

($s_{i,c}$). Assuming cellular physiology changes as growth slows, the growth rate was used to find intermediate mRNA concentrations.

$$\begin{aligned} M_{i,c}(t) &= \frac{N_{NT,Control}(0) \cdot s_{i,Control}}{\sum_j L_j \cdot s_{j,Control}} \cdot \frac{\mu_c(3.5) - \mu_c(t)}{\mu_c(3.5) - \mu_c(0)} \\ &\quad + \frac{N_{NT,c}(3.5) \cdot s_{i,c}}{\sum_j L_j \cdot s_{j,c}} \cdot \frac{\mu_c(t) - \mu_c(0)}{\mu_c(3.5) - \mu_c(0)} \end{aligned} \tag{4.4}$$

After finding the individual mRNA concentrations, the cell-wide protein synthesis

rate was determined from the number of active protein producing 70S ribosomes (N_{70S}). The initial number of ribosomes ($N_R(0) = 18.6 \times 10^3$ ribosomes per cell) was found by interpolating the experimentally observations in Table 3 of Bremer and Dennis [13] at a doubling rate of 1.2 hr^{-1}. The fraction of these ribosomes actively producing protein ($\beta_{R,c}$) was assumed to scale with the growth rate between 80% [13] at induction and the active fraction (70S and polysomes) observed 3.5 hours afterwards from ribosome elution data [2]:

$$\beta_{R,c}(t) = 0.8 \cdot \frac{\mu_c\,(3.5) - \mu_c\,(t)}{\mu_c\,(3.5) - \mu_c\,(0)} + \frac{A_{70S,c} + A_{Poly,c}}{A_{70S,c} + A_{100S,c} + A_{Poly,c}} \cdot \frac{\mu_c\,(t) - \mu_c\,(0)}{\mu_c\,(3.5) - \mu_c\,(0)} \tag{4.5}$$

where $A_{70S,c}$, $A_{100S,c}$, $A_{Poly,c}$ are the areas of the 70S, 100S, and polysome peaks respectively from a fit of the ribosome elution data collected at 3.5 hours after induction for each culture condition to a series Cauchy-Lorentz-distributed peaks (Figure 4.3) (see Supplementary Material). With no accurate estimate of ribosome turnover rates, we assumed net ribosome synthesis halts following induction because the growth rate appears to slow immediately. Thus the only reduction in ribosomes is dilution by growth, whereby the number of actively translating ribosomes was defined as:

$$N_{70S,c}(t) = N_{R,0} \cdot \beta_{R,c}(t) \cdot \frac{X_c(0)}{X_c(t)} \tag{4.6}$$

4.2.4 Translation Rates

With the translation substrates defined, protein synthesis was coarse-grained into three distinct steps: initiation, elongation, and termination. The initiation rate is proportional to the translation initiation efficiency calculated using the structured ribosome binding site model of de Smit and van Duin [19]. This model assumes the diffusion of the bulky 30S ribosome subunit, not 30S-mRNA binding, is rate-limiting. The only adjustable parameters were the 30S ribosome concentration in each culture condition, defined as one

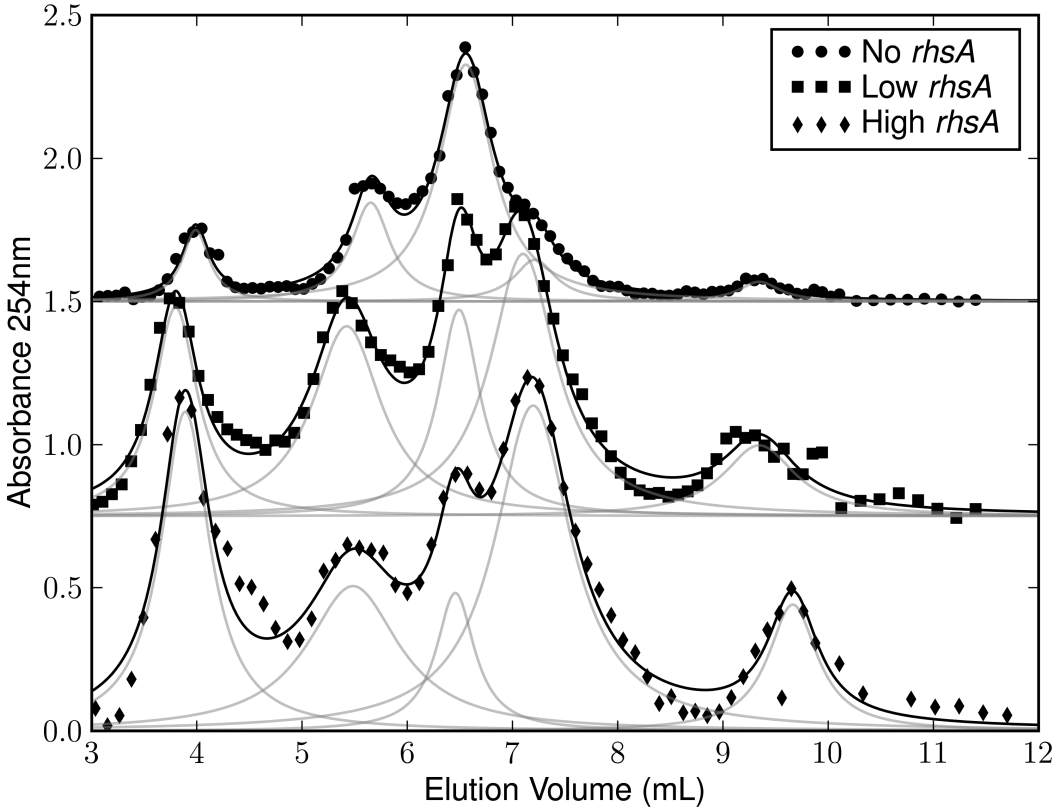

Figure 4.3: Ribosome elution absorbance at 254nm (points) [2] was fit using the sum of Cauchy-Lorentz-distributed peaks (black line) representing the 30S, 50S, 70S, 100S, and polysome ribosome fractions (gray lines from left to right, respectively). See the Supplementary Material for a full description and peak parameters. For clarity, Low and No were shifted up 0.75 and 1.5, respectively.

quarter of the active 70S ribosomes in Equation 4.6, and the free energy of unfolding the ribosome binding site found using the UNAFold software suite [56]. The resulting efficiency was constrained at 10% or higher to prevent genes with artificially low translation initiation rates. These translation initiation efficiencies were linearly scaled as described below to satisfy the expected concentration of 70S ribosomes. A full description of translation initiation kinetics and associated parameter values is provided in the Supplementary Material.

Translation elongation rates are taken from the work of Fluitt et al. [21]. To describe the expected slowing of translating ribosomes as the cell culture enters stationary phase, the individual rates were scaled by the change in experimentally observed peptide chain elongation rates $c_{p,c}(t)/c_{p,c}(0)$, where $c_{p,c}$ is a linear fit of the data presented by Bremer and Dennis [13], defined in Figure 4.2. Finally, translation termination was assumed not rate-limiting and set to a rate of 10^6 min^{-1}.

4.2.5 Mathematical Model

Using the above description of mRNA concentrations, 70S ribosomes, and translation kinetics, the specific rate of protein synthesis ($J_{i,c}$) may be fixed by finding the maximum translation initiation rate which satisfies the expected number of active 70S ribosomes:

$$N_{70S,c}(t) = \sum_{i}^{genes} \left(M_{i,c}(t) \sum_{j}^{codons} \rho_{i,c,j}(t) \right) \tag{4.7}$$

where $\rho_{i,c,j}$ is the probability of finding a ribosome at codon j of gene i in condition c found from a stochastic simulation of the lattice gas model of translation [80]. Stochastic simulations are used because deterministic, or mean-field, solutions encounter numerical difficulties leading to inaccurate estimates of ρ [35, 79]. Simulations of 128,000 proteins synthesized per gene provided sufficiently accurate estimates of $\rho_{i,c,j}$ and $J_{i,c}$, with an average relative error of 0.3% in the polysome size and specific protein synthesis rate. The specific protein production rates were found at half-hour increments from induction to 3.5 hours and interpolated using a third order β-spline.

The individual protein concentrations were found by integrating:

$$\frac{\mathrm{d}\, P_{i,c}}{\mathrm{dt}} = J_{i,c}(t) \cdot M_{i,c}(t) - k_{d,P} \cdot P_{i,c}(t) \tag{4.8}$$

$$P_{i,c}(t=0) = \frac{J_{i,c}(t=0) \cdot M_{i,c}(t=0)}{\mu_c(t=0) \cdot \ln 2} \tag{4.9}$$

The primary mechanism of protein degradation was assumed to be dilution by growth, thus the degradation rate $k_{d,P}$ was set to the growth rate: $\mu_c(t) \cdot \ln 2$. The initial concentration of protein at induction was assumed to be at steady-state.

To understand the dynamics of protein fold changes, the standard method of plotting changes in transcriptomic and proteomic observations, Equation 4.8 may be expressed by dividing by the initial protein concentration (Equation 4.9).

$$\frac{1}{P_{i,c}(t=0)} \cdot \frac{\mathrm{d}\, P_{i,c}}{\mathrm{dt}} = \frac{\mathrm{d}\, f_{i,c}^{P}}{\mathrm{dt}} = \mu_c(t=0) \cdot \ln 2 \cdot f_{i,c}^{J}(t) \cdot f_{i,c}^{M}(t) - \mu_c(t) \cdot f_{i,c}^{P}(t) \tag{4.10}$$

where $f_{i,c}^{J}$, $f_{i,c}^{M}$, and $f_{i,c}^{P}$ are the fold changes of specific protein production rate, mRNA concentration, and protein concentration relative to their initial values, respectively. In a simple conceptual experiment of instant changes in growth rate, mRNA, and specific protein production rate, Equation 4.10 may be exactly solved:

$$f_{i,c}^{P}(\infty) = \frac{\mu_c(t=0) \cdot \ln 2}{\mu_c} \cdot f_{i,c}^{J} \cdot f_{i,c}^{M} \tag{4.11}$$

$$f_{i,c}^{P}(t) = f_{i,c}^{P}(\infty) + \left(1 - f_{i,c}^{P}(\infty)\right) \cdot \exp\left(\mu_c \cdot t\right) \tag{4.12}$$

4.2.6 χ^2 Quantification

The model results are compared against measured protein concentrations from a previously described iTRAQ$^{\mathrm{TM}}$ shotgun proteomics experiment [1]. Briefly, the mean and standard deviation of the mass spectrum reporter ion peak areas of each peptide were found for each protein with two or more peptides. The minimum standard deviation set at 20% of the mean to prevent an artificially high confidence resulting from proteins with few peptide measurements and thus a poor estimate of the standard deviation. The resulting dataset contained 525 proteins at each data point. The agreement between the

model results and the experimental data was quantified using a least squares function:

$$\chi^2 = \sum_{c}^{expts} \chi_c^2 = \sum_{c}^{expts} \sum_{i}^{genes} \left(\frac{\beta_c \cdot \gamma_i \cdot P_{i,c} - P_{i,c}^{obs}}{\sigma_{i,c}} \right)^2 \tag{4.13}$$

Two separate types of scaling factors facilitated a comparison between predicted protein concentrations and measured mass spectrum peak areas. β_c captured the average total protein per cell and was necessary because identical amounts of protein were loaded from each culture condition. γ_i was the proportionality constant of mass spectrum peak area to protein concentration for gene i in all data points. These scaling factors were found by iteratively solving $\frac{\partial \chi^2}{\partial \beta_c} = 0$ and $\frac{\partial \chi^2}{\partial \gamma_i} = 0$.

4.3 Results

4.3.1 Model Validation

The model results are validated in two ways. First, the total protein concentration is compared to experimental observations to verify the balance between protein synthesis and degradation. As shown in Figure 4.4, the model results show the total protein concentration falling from 15×10^8 to 10×10^8 amino acids per cell. The slight increase at later times for No and High is the result of the protein degradation falling faster than the decreasing rate of protein synthesis. Experimentally, the total protein concentration is observed at 10.4×10^8 amino acids per cell at a stable doubling rate of 1.2 hr^{-1} [13]. While slightly higher, the total protein levels are consistent with those observed experimentally reflecting a biologically reasonable balance of protein synthesis and degradation.

Next the agreement between the model results and the experimental observations of individual protein concentrations is calculated using the χ^2 function (defined in Meth-

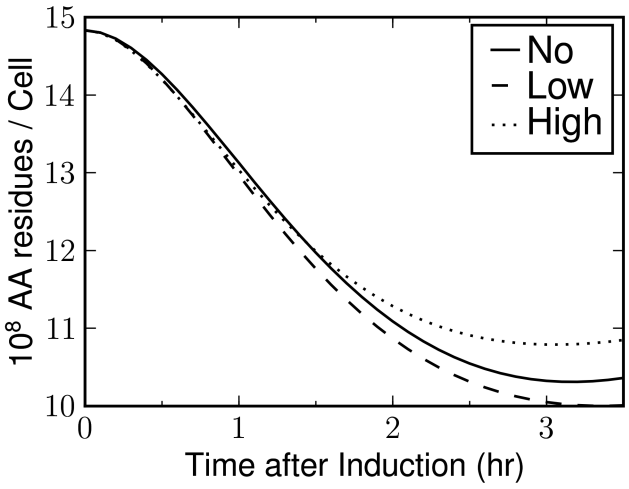

Figure 4.4: Predicted total protein concentration defined as the number of amino acids (AA) per cell.

ods). The results for each data point are $\chi^2_{control} = 355$, $\chi^2_{no} = 403$, $\chi^2_{low} = 265$, and $\chi^2_{high} = 266$ for a total $\chi^2 = 1289$. With 525 genes measured at 4 data points less $525 + 4$ scale factors, this χ^2 value is significantly less than the 1571 degrees of freedom. Thus, the model results are consistent with the experimental observations and may present an accurate description of the proteome.

4.3.2 Predicted Proteome

To understand the predicted proteome presented below, we first consider the conceptual experiment of instantaneous changes in mRNA, specific protein production rate, and growth rate, as presented in Figure 4.5. The evolution of the protein fold change is shown from no initial change to nearly steady state after 10 doublings. The lower protein fold changes are observed to equilibrate much slower than higher protein fold changes, a result of the exponential growth or decay which occurs in Cartesian space. That is, the exponential growth or decay dictates halving the difference between the current and

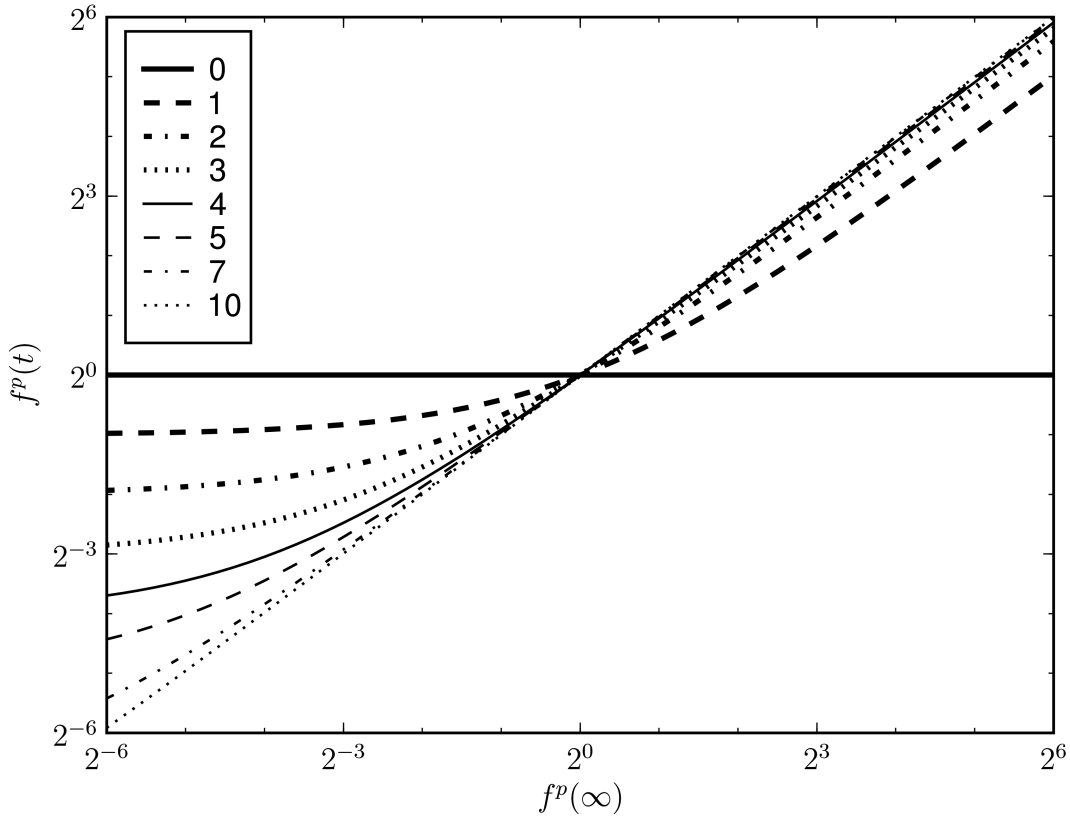

Figure 4.5: Evolution of protein fold changes ($f^P(t)$) to an instantaneous cellular change ($f^P(\infty)$), as defined in Equations 4.11 and 4.12. The legend defines the number of doublings since the instantaneous cellular perturbation for each line.

infinite fold change (1:1 line) every doubling, which when plotted in log-log space skews the response of low versus high protein fold changes.

A similar overall trend is observed in the predicted proteomes shown in upper panels of Figure 4.6. As suggested by the growth data, fewer than three doublings occur during the 3.5 hours following induction, resulting in a dynamically changing proteome yet to achieve steady state. The main body of points, following a similar trend to that observed in Figure 4.5, results from the majority of genes sharing a common specific protein production rate. The spread observed downward and to the right of this main body is a

result of genes with relatively slow translation rates, highlighted by the red points which have the minimum allowable translation initiation efficiency.

As suggested by the χ^2 value presented above, the model results are consistent with the experimental observations, shown in the lower panels of Figure 4.6. The slight increasing trend present in No:Control compared to the other fold changes is consistent with the observed higher growth rate. But the experimental data has a considerably wider distribution than the model results. This difference may be a result of inherent inaccuracies in the measurement techniques and / or gene regulation not captured by the model. The majority of poorly fit genes fall either above and left or below and right of the general body of points. As mentioned above, spread below and to the right may be the result of lowered specific protein synthesis rates, or targeted protein degradation. Conversely, points above and left may be the result of higher specific translation rates (i.e. preferentially better ribosome recruitment), such as that expected from a transcript with an activated riboswitch.

4.3.3 Functional Classification

Protein synthesis and degradation are highly complex processes which include gene-specific levels of regulation that may not be captured by the model results. Some examples include targeted protein degradation, riboswitches which modulate translation, and possible folding of mRNA or binding of factors which inhibit translation elongation. Thus, poorly fit genes may contain post-transcriptional regulation not captured by the model description.

To investigate any possible post-transcriptional regulation, the 60 worst-fit genes are clustered to highlight any commonly repressed or activated pathways, shown in Fig-

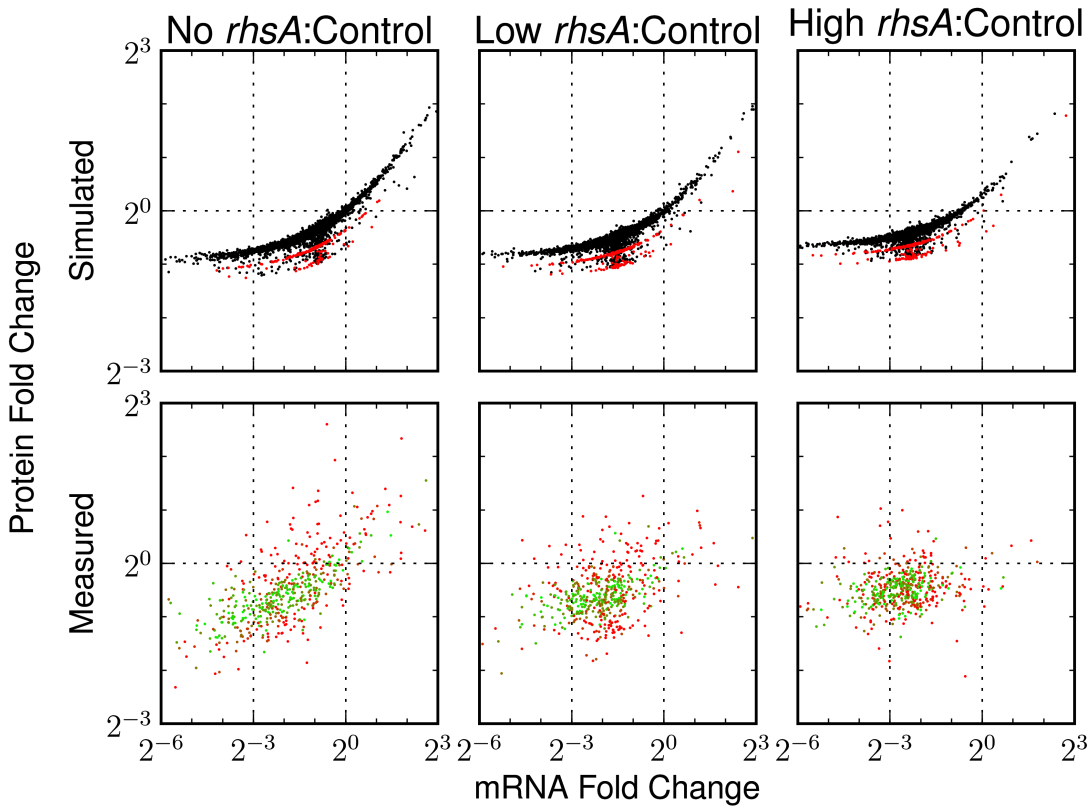

Figure 4.6: Predicted (top panels) and experimentally observed (bottom panels) mRNA vs. protein fold changes for No, Low, and High with respect to Control. The red points shown in the top panels have the lowest allowable translation initiation efficiency. The experimentally observed protein fold changes are scaled by the scaling factor $\beta_{Control}/\beta_c$ to account for differences in protein loading for the proteomic measurements. These values are colored from green (perfectly fit) to red (poorly fit).

ure 4.7. The color represents the $\chi_{i,c}$ value, where green represents relatively higher than expected and red relatively lower than expected protein concentrations compared to the same gene in other experiments (i.e. green does not necessarily indicate a high protein fold change but a higher than expected protein fold change). Four groups of poorly fit genes emerge: genes which are preferentially repressed in the Control, No, Low, and High datasets colored with green, red, purple, and blue names respectively. The Control

cluster (green names) suggests all cultures specifically up-regulate gluconeogenesis and pyrimidine synthesis, the former of which is expected as a result of starvation in stationary phase. Interestingly, the No dataset additionally activates purine synthesis, which is increasingly lacking with increasing *rhsA* expression, shown in the High cluster (blue names). Additionally, the *yfiA* gene (a putative 70S dimer inhibitor) is activated in Low and High relative to No possibly underlying the abnormally high 100S concentration observed.

Several of these genes exhibited possible riboswitch control in other bacteria [54], including the *gcvTHP* operon in *Bacillus subtilis* [55] and *mgtA* in *Salmonella* [85]. Additionally, the clustered presence of nearly all the pyrimidine ribonucleotide synthesis enzymes (6 of 9 observed proteins) and purine ribonucleotide synthesis enzymes (5 of 6 observed proteins in the upper shunt) are not consistent with being the result of random experimental error. The presence of both riboswitches and nearly complete metabolic pathways may be consistent with post-transcriptional regulation not captured by the model.

4.4 Discussion

The mathematical model of protein production presented here is consistent with experimental observations of individual and cell-wide protein concentrations. The nonlinear relationship predicted by the model results highlights the dynamic changes present in the proteome, even after several hours of induction. As presented in Figure 4.5, a minimum of seven to ten divisions (5 to 10 hours) would be required for the proteome to achieve steady state relative to the transcriptome under the ideal situation of instantaneous cellular changes. These results reinforce the need to explicitly measure protein

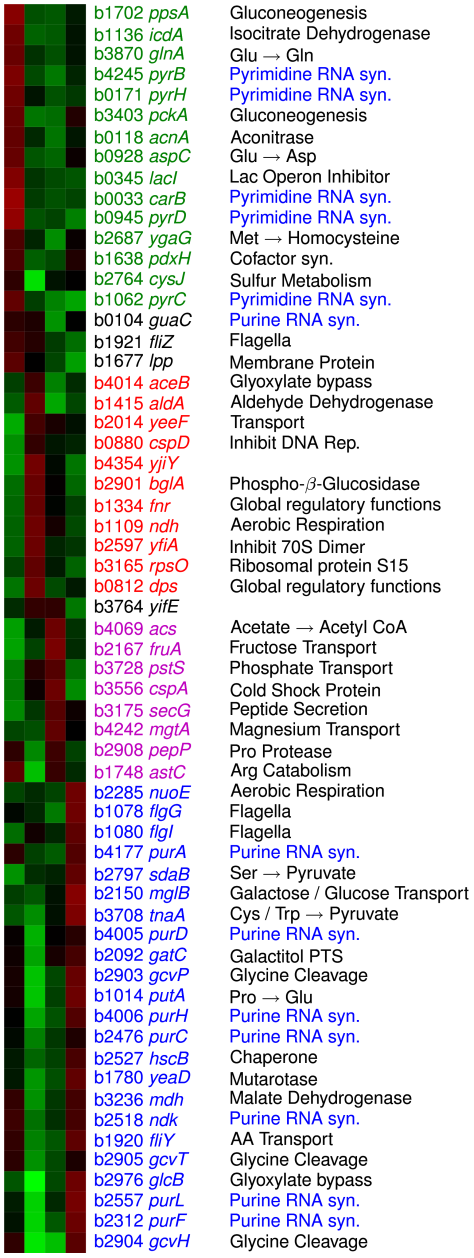

Figure 4.7: Hierarchical clustering of the 60 most poorly described genes. The color represents the $\chi_{i,c}$ value where red shows lower than expected protein levels and green represents higher than expected levels compared to the same gene in other conditions. Gene blattner numbers and names are colored to reflect preferentially repressed genes in the Control (green), No (red), Low (purple), or High (blue) data points. Gene functions (or putative functions) are also shown with pyrimidine and purine synthesis genes highlighted in blue.

concentrations, or to more carefully address any assumed transcriptome-proteome relationship, when studying cellular perturbations.

Interestingly, the genes poorly described by the model results fall into several distinct metabolic pathways. Comparing the results of increasing levels of *rhsA* induction illustrates the relative decrease in purine synthesis and inhibition of 100S ribosome formation. These results are consistent with the observed slowing of growth, possibly because of unbalanced metabolism, and the response to abnormally high 100S ribosome concentration. Additionally, the poor fitting of nearly complete metabolic pathways and the presence of riboswitch-regulated genes is consistent with post-transcriptional regulation not captured by the model representation. Thus, beyond the heterogeneity of ribosome binding sites and codon bias, translation regulators including riboswitches represent an opportunity for future model development. One such post-transcriptional regulation mechanism not specifically mentioned above, but implicit in the definition of Equation 4.8, is targeted protein degradation. If protein degradation is significant relative to dilution by growth, the protein fold change of this protein would more quickly acheive steady state with respect to changes in mRNA. Particularly interesting is the likely increase the proteome distribution and consistency with experimental observations resulting from the additional regulation. More experimental evidence and better bioinformatic prediction capabilities are critical to facilitating this model development.

Here we present a cell-wide model of protein synthesis capable of predicting the proteome composition based upon the relatively easily measured transcriptome. The predictions capture the majority of experimentally measured protein concentration fold changes, with outlying genes suggesting post-transcriptional regulation such as riboswitches. Incorporating this level of regulation into future models could provide an even more accurate representation of the proteome and improve our understanding of

translation regulation.

4.5 Acknowledgments

We would like to thank Kunal Aggarwal for collecting the experimental observations and useful discussions regarding the resulting data normalization.

4.6 Supplementary Material

4.6.1 Ribosome Elution Data Peak Areas

Ribosome elution data was preconditioned with a straight-line filter and fit to a series of Cauchy-Lorentz peaks using OriginPro v7.5 (Northampton, MA). The probability density function for a Cauchy-Lorentz distribution located at x_0 with a scale parameter of γ is:

$$f(x; x_0, \gamma) = \frac{1}{\pi} \cdot \frac{\gamma}{(x - x_0)^2 + \gamma^2} \tag{4.14}$$

The plotted peaks (gray lines in Figure 4.3) are the product of the fit peak area (A) and the probability density function. Letting the scale parameter, $\gamma = \frac{A}{\pi M}$, where M is the peak maximum gives:

$$A \cdot f(x; x_0, A, M) = \frac{A}{\pi} \cdot \frac{\frac{A}{\pi M}}{(x - x_0)^2 + \left(\frac{A}{\pi M}\right)^2} \tag{4.15}$$

The parameters found from fitting the location (x_0), peak area, and maximum are shown in Table 4.1 below.

Table 4.1: Cauchy-Lorentz distribution parameters fit to the ribosome elution data

Condition	Peak	Location	Maximum	Area
No	30S	3.988392	0.249731	0.120562
	50S	5.652212	0.344910	0.245903
	70S	6.556273	0.827115	0.803320
	100S	7.211697	0.145121	0.110101
	Polysome	9.344015	0.068411	0.050908
Low	30S	3.797691	0.729632	0.573751
	50S	5.424066	0.663189	0.831335
	70S	6.492035	0.720899	0.548695
	100S	7.100364	0.915994	1.099713
	Polysome	9.345341	0.247154	0.325850
High	30S	3.889095	1.114605	0.986692
	50S	5.483799	0.505572	0.850137
	70S	6.456521	0.481782	0.336935
	100S	7.195709	1.136296	1.569002
	Polysome	9.663747	0.441150	0.395809

4.6.2 Translation Initiation Efficiency

To clarify several typographic errors in the derivation and parameters presented by de Smit and van Duin [19], the following explicitly states the equation for the translation initiation efficiency (E) and the parameters used (shown in Table 4.2). Please see the author's work for a more complete description of the model and parameters.

The author's model represents the ribosome binding site as unfolded ($[U]$) or folded ($[F]$). The 30S ribosome can associate with either state ($[30S \cdot U]$ and $[30S \cdot F]$), of

which only the 30S bound to the unfolded ribosome binding site can form the initiation complex ($[IC]$) and initiate translation. The translation initiation efficiency (E) is described as the concentration of ribosomes in the initiation complex relative to the other intermediate states. This solution was found using Mathematica v6 (Wolfram Research, Inc., Champaign, IL).

$$\begin{aligned}
[U] \cdot D &= k_{-F} \cdot k_{-SF} \cdot \{k_i \cdot k_c + (k_c + k_{-i})(k_{-SU} + k_{-SFU})\} \\
&\quad + k_{SFU} \cdot (k_{-F} + k_{SF}[30S]) \{k_i \cdot k_c + (k_c + k_{-i}) \cdot k_{-SU}\} \quad (4.16) \\
[F] \cdot D &= k_{-SF} \cdot k_{-SFU} \cdot (k_c + k_{-i})(k_F + k_{SU}[30S]) \\
&\quad + k_F \cdot (k_{-SF} + k_{SFU}) \{k_i \cdot k_c + (k_c + k_{-i}) \cdot k_{-SU}\} \quad (4.17) \\
[30S \cdot F] \cdot D &= k_F \cdot k_{SF}[30S] \{k_i \cdot k_c + (k_c + k_{-i})(k_{-SU} + k_{-SFU})\} \\
&\quad + k_{SU}[30S] k_{-SFU} \cdot (k_c + k_{-i})(k_{-F} + k_{SF}[30S]) \quad (4.18) \\
[30S \cdot U] \cdot D &= k_{SF}[30S] k_{SFU} \cdot (k_c + k_{-i})(k_F + k_{SU}[30S]) \\
&\quad + k_{SU}[30S] k_{-F} \cdot (k_c + k_{-i})(k_{-SF} + k_{SFU}) \quad (4.19) \\
[IC] \cdot D &= k_{SF}[30S] k_{SFU} \cdot k_i \cdot (k_F + k_{SU}[30S]) \\
&\quad + k_{SU}[30S] k_{-F} \cdot k_i \cdot (k_{-SF} + k_{SFU}) \quad (4.20) \\
E &= \frac{[IC]}{[U] + [F] + [30S \cdot F] + [30S \cdot U] + [IC]} \quad (4.21)
\end{aligned}$$

$[30S]$ and ΔG^0_f (defined below) are the only independent variables for each culture condition and gene, respectively. The remaining parameters are taken directly for de Smit and van Duin [19], with the exception of k_{-SF} whose value was not given but found via fitting the data the authors presented.

The concentration of active 30S ribosomes (in molar, not number per cell) was assumed to be a quarter of the actively translating ribosomes (from the observation that 80% of ribosomes are translating during exponential growth [13]):

$$[30S] = \frac{N_{70S,c}(t)}{4 \cdot N_A \cdot (10^{-15}\ \mathrm{L})} \quad (4.22)$$

Table 4.2: Translation Initiation Parameters

Name	Value	Description
k_F	1.5×10^7 s^{-1}	Ribosome binding site folding
k_{-F}	$k_F / \exp\left(-\frac{\Delta G_f^0}{R \cdot T}\right)$	Ribosome binding site unfolding
k_{SU}	10^7 M^{-1} s^{-1}	Diffusion-limited unfolded binding
k_{-SU}	1 s^{-1}	30S unbinding from unfolded mRNA
k_{SF}	k_{SU}	Diffusion-limited folded binding
k_{-SF}	30.0 s^{-1}	30S unbinding from folded mRNA
k_{SFU}	k_{-F}	mRNA unfolding with bound 30S
k_{-SFU}	$\frac{k_F \cdot k_{-SU} \cdot k_{SFU} \cdot k_{SF}}{k_{-F} \cdot k_{SU} \cdot k_{-SF}}$	mRNA folding with bound 30S
k_i	10^5 s^{-1}	Initiation complex formation
k_{-i}	0.0	Non-reversible initiation complex formation
k_c	0.5 s^{-1}	Ribosome binding site clearance

where $N_{70S,c}(t)$ is defined in the text as Equation 4.6, N_A is Avogadro's number, and the cellular volume is assumed to be 10^{-15} L.

The free energies of SD folding / unfolding were found using the `hybrid-min` and `hybrid-ss-min` programs of the UNAFold software suite [56]. `hybrid-min` was used to search for the Shine-Dalgarno (SD) mRNA-rRNA interaction of the ribosome binding site. The SD-binding sequence of the 16S rRNA was represented by `5'-gaucAccuccuua-3'` where the capital `A` was aligned to the -20 to -3 regions upstream of each gene using the default program options except disallowing bulges, allowing single isolated basepairs, and fixing the alignment of the two sequences. The lowest free energy alignment was assumed to be the ribosome binding site. Next the entire mRNA with 50 additional upstream nucleotides was allowed to fold using `hybrid-ss-min` with the default options and either allowing or disallowing basepairs

to the identified ribosome binding site. The difference between these two states yielded the free energy of folding / unfolding the ribosome binding site (ΔG_f^0) used to find k_{-F}.

CHAPTER 5

IS CODON BIAS SUFFICIENT TO PREDICT THE BACTERIAL PROTEOME?

As new and more complex models of protein synthesis are presented in the literature, it is important to assess whether the biologically relevant regulation is captured. One common level of regulation included in these models is the either homogeneous or inhomogeneous kinetics of translation elongation, which can be interpreted as a measure of codon bias. Using experimental observations of the transcriptome and proteome, we investigate whether heterogeneity in translation elongation can produce the expected changes in protein synthesis rates. Our results are consistent with regulation based purely on codon bias being insufficient to describe the observed global changes in protein synthesis.

5.1 Introduction

Protein synthesis is a fundamental cellular process, producing much of the cell's functional content from enzymes to regulators. While measuring protein expression is important to determine the cellular state, the current proteomic measurement techniques are expensive in time and resources and lack the genome-wide coverage of transcriptomic measurements. As a result transcript changes are commonly assumed either implicitly or explicitly to reflect changes in the proteome, an assumption which appears to contradict the complex and nonlinear changes in mRNA and protein observed across several organisms including *Escherichia coli* [48, 49], yeast [31, 38, 71], and the human liver [4].

In an attempt to describe the complex process of protein synthesis, numerous math-

ematical models have been developed [24, 80, 100]. These models incorporate diverse levels of complexity: from a detailed translation elongation description [100] to coarse-grained translation elongation [24, 80] to other translation regulators such as ribosome recycling [24]. These models are primarily applied to the prediction of the specific protein production rate of an individual gene, with relatively few works extending the model description to cell-wide protein synthesis. One cell-wide study of protein synthesis is that of Mehra and Hatzimanikatis [60], although their results were based upon a hypothetical *E. coli* transcriptome.

Here we present a comparison of the experimentally observed changes in protein synthesis rates to the predicted changes from a cell-wide model of protein synthesis. The experimental results highlight the expected change in protein synthesis rates, spread over two- to three- fold, which appear to be the result of the underlying biological regulation and not experimental uncertainty. Similar to previous work [60], the model presented here captures the competition for cellular resources (i.e. ribosomes) by the experimentally observed *E. coli* transcriptome. Comparatively, the model results for several sets of inhomogenous translation kinetics predict a significantly more modest change in protein synthesis rates. The results of this comparison are consistent with gene-specific regulation beyond codon bias may being primarily responsible for the experimentally observed change in protein synthesis and therefore also primarily responsible for the complex and nonlinear mRNA-protein relationship.

5.2 Methods

5.2.1 Reduction of Experimental Data

mRNA and protein fold changes are taken from *rhsA* over-expression in *E. coli* [1, 2]. Here we focus on an individual experiment, the case of no *rhsA* induction, with other conditions producing similar results. mRNA and protein concentrations were measured using microarray [2] and shotgun proteomic [1] measurements, respectively, immediately prior to (no) induction and 3.5 hours afterwards. To facilitate a more direct comparison to the model results, the expected changes in specific protein synthesis rates per mRNA are derived as follows. First, a simplified view of protein synthesis was employed:

$$\frac{\mathrm{d}\,P_i}{\mathrm{dt}} = J_i(t) \cdot M_i(t) - k_{d,P}(t) \cdot P_i(t) \tag{5.1}$$

$$P_i(t=0) = \frac{J_i(t=0) \cdot M_i(t=0)}{k_{d,P}(t=0)} \tag{5.2}$$

where for gene i, P_i and M_i are the protein and mRNA concentrations, respectively, J_i is the specific protein synthesis rate, and $k_{d,P}$ is the protein degradation rate. The initial protein concentration in Equation 5.2 is found assuming the steady state at $t = 0$. Assuming protein degradation is primarily a result of dilution by growth, $k_{d,P}$ is set equal to the growth rate $\mu(t)$:

$$\mu\,(t) = \frac{k \cdot \left(1 - \frac{X_0}{X_\infty}\right)}{1 - \frac{X_0}{X_\infty}\left(1 - \exp\left(k \cdot t\right)\right)} \tag{5.3}$$

The parameters were fit to experimentally observed growth data [2], finding initial and infinite biomass concentrations of $X_0 = 0.45$ and $X_\infty = 2.92$, respectively, and a maximum growth rate of $k = 0.95$ hr^{-1}.

To find the evolution of fold changes with respect to time, Equation 5.1 is divided

by the initial concentration:

$$\frac{1}{P_i(t=0)} \cdot \frac{\mathrm{d}\, P_i}{\mathrm{dt}} = \frac{\mathrm{d}\, f_i^P}{\mathrm{dt}} = \mu(t=0) \cdot f_i^J(t) \cdot f_i^M(t) - \mu(t) \cdot f_i^P(t) \tag{5.4}$$

where f_i^J, f_i^M, and f_i^P are the fold changes of specific protein production rate, mRNA concentration, and protein concentration relative to their initial values, respectively. If changes in mRNA and the specific protein synthesis rate (but not growth rate) are assumed to occur instantaneously, $f_i^P(t)$ may be exactly solved:

$$f_i^P(t) = \frac{f_i^J \cdot f_i^M}{\mu(t)} \cdot \exp\left[-k \cdot t\right] \cdot \left(1 + \frac{\mu(0)^2 \cdot \ln\left[1 - \frac{X_0}{X_\infty}\left(1 - \exp\left[k \cdot t\right]\right)\right]}{k \cdot \frac{X_0}{X_\infty}}\right) \tag{5.5}$$

which may be solved for the expected change in specific protein synthesis rate (f^J), given the observed changes in the mRNA and protein concentrations mentioned above.

As a result of experimental protocols and data normalization in the measured changes in mRNA and protein, the absolute value of f^M and f^P measured experimentally must be scaled by one factor for all genes. Thus the mean expected change in specific protein synthesis rate presented below was scaled by the geometric mean to facilitate comparison to model results.

5.2.2 Translation Substrates

mRNA expression data were collected as described in [2]. Briefly, the mean signal intensity of each gene was found from four replicates (two biological, two technical) at each experimental data point. After excluding open reading frames with in-frame stop codons or frameshifting, a set of 4117 open reading frames represented the *in silico* transcriptome. To scale the signal intensities to actual concentrations, the total mRNA nucleotide concentration at each experimental condition (N_{NT}) was found from the experimentally

observed nucleotide synthesis rate (r_m), a function of the growth rate, subject to a constant degradation rate ($k_{d,m} = 1min^{-1}$) [13]:

$$\frac{dN_{NT}}{dt} = r_m(t) - k_{d,m} \cdot N_{NT}(t) = 9.25 \times 10^5 \cdot \frac{\mu(t)}{\ln 2} - k_{d,m} \cdot N_{NT}(t) \tag{5.6}$$

The nucleotide synthesis rate (r_m) can be linearly fit to the data given at different growth rates in Table 3 of Bremer and Dennis [13]. The calculated total nucleotide concentrations are 10.7×10^5 at t=0 and 4.34×10^5 3.5 hrs afterwards. The individual mRNA concentrations at each experimental data point were calculated from the total nucleotide concentration, the gene lengths (L_i), and microarray signal intensities (s_i).

$$M_i(t) = \frac{N_{NT}(t) \cdot s_i}{\sum_j L_j \cdot s_j} \tag{5.7}$$

After finding the individual mRNA concentrations, the cell-wide protein synthesis rate was determined from the number of active protein producing 70S ribosomes (N_{70S}). The initial number of ribosomes ($N_R(0) = 18.6 \times 10^3$ ribosomes per cell) was found by interpolating the experimentally observations in Table 3 of Bremer and Dennis [13] at a doubling rate of 1.2 hr^{-1}. The fraction of these ribosomes actively producing protein (β_R) was assumed to be 80% at induction [13] and 89% as experimentally determined 3.5 hours later [2]. With no accurate estimate of ribosome turnover rates, we assumed the net ribosome synthesis halts following induction because the growth rate appears to slow immediately. Thus the only reduction in ribosomes is dilution by growth, whereby the number of actively translating ribosomes was defined as:

$$N_{70S}(t) = N_R(0) \cdot \beta_R \cdot \frac{X(0)}{X(t)} \tag{5.8}$$

where $N_{70S}(t = 0) = 13907$ and $N_{70S}(t = 3.5) = 2875$.

5.2.3 Translation Model

With the translation substrates defined, protein synthesis was coarse-grained into three distinct steps: initiation, elongation, and termination. The initiation rate is proportional to the translation initiation efficiency calculated using the structured ribosome binding site model of de Smit and van Duin [19]. This translation initiation model assumes the diffusion of the bulky 30S ribosome subunit, not 30S-mRNA binding, is rate-limiting. The only adjustable parameters used were the 30S ribosome concentration in each culture condition, defined as one quarter of the active 70S ribosomes defined in Equation 5.8, and the free energy of unfolding the ribosome binding site found using the UNAFold software suite [56]. The translation initiation efficiencies calculated were linearly scaled as described below to satisfy the expected concentration of 70S ribosomes.

Translation elongation rates were taken from the work of Fluitt et al. [21]. To describe the expected slowing of translating ribosomes as the cell culture enters stationary phase, the individual rates were scaled by approximately 75% at 3.5 hours as experimentally observed [13]. Translation termination was assumed not rate-limiting and set to a rate of 10^6 min^{-1}.

Using the above description of mRNA concentrations, 70S ribosomes, and translation kinetics, the specific rate of protein synthesis (J_i) was fixed by finding the maximum translation initiation rate which satisfies the expected number of active 70S ribosomes:

$$N_{70S}\left(t\right) = \sum_{i}^{genes}\left(M_i\left(t\right)\sum_{j}^{codons}\rho_{i,j}\left(t\right)\right) \tag{5.9}$$

where $\rho_{i,j}$ is the probability of finding a ribosome at codon j of gene i found from a stochastic simulation of the lattice gas model of translation [80]. Stochastic simulations are used because deterministic, or mean-field, solutions encounter numerical difficulties leading to inaccurate estimates of ρ [35, 79]. Simulations of 128,000 proteins synthe-

sized per gene provided sufficiently accurate estimates of $\rho_{i,j}$ and J_i, with an average relative error of 0.3% in the polysome size and specific protein synthesis rate. The change in specific protein synthesis rate can thus be found as $f_i^J = J_i(3.5)/J_i(0)$. As mentioned in the previous section, the experimental techniques require a single scaling factor to compare the experimental observations to model results. Thus the distribution of changes in protein synthesis rates, presented below in Figure 5.3, is similarly scaled by the geometric mean to facilitate comparison.

5.3 Results

5.3.1 Experimentally Observed Synthesis Rates

The experimentally observed changes in specific protein synthesis rates (f_i^J), described by Equation 5.5 developed in the Methods, are shown in the left panel of Figure 5.1. The observed distribution in the experimental observations may be the result of inaccurate measurements or the underlying regulation of protein concentration. To differentiate these two possibilities, the experimental data is split into two groups: 284 genes with 5 or fewer peptides (low confidence means) and 241 proteins with greater than 5 peptides (high confidence means). If the distribution width is a result of protein measurement uncertainty, the high confidence means would be expected to have a narrower distribution than the low confidence means. This test assumes the majority of measurement uncertainty is from the protein measurements and there is no systematic error (not regulation) specific to a protein and cell culture condition, which would not be captured by this comparison. A statistical measure of the variance difference is not presented because the expected distribution of either the experimental error (the ratio of protein and

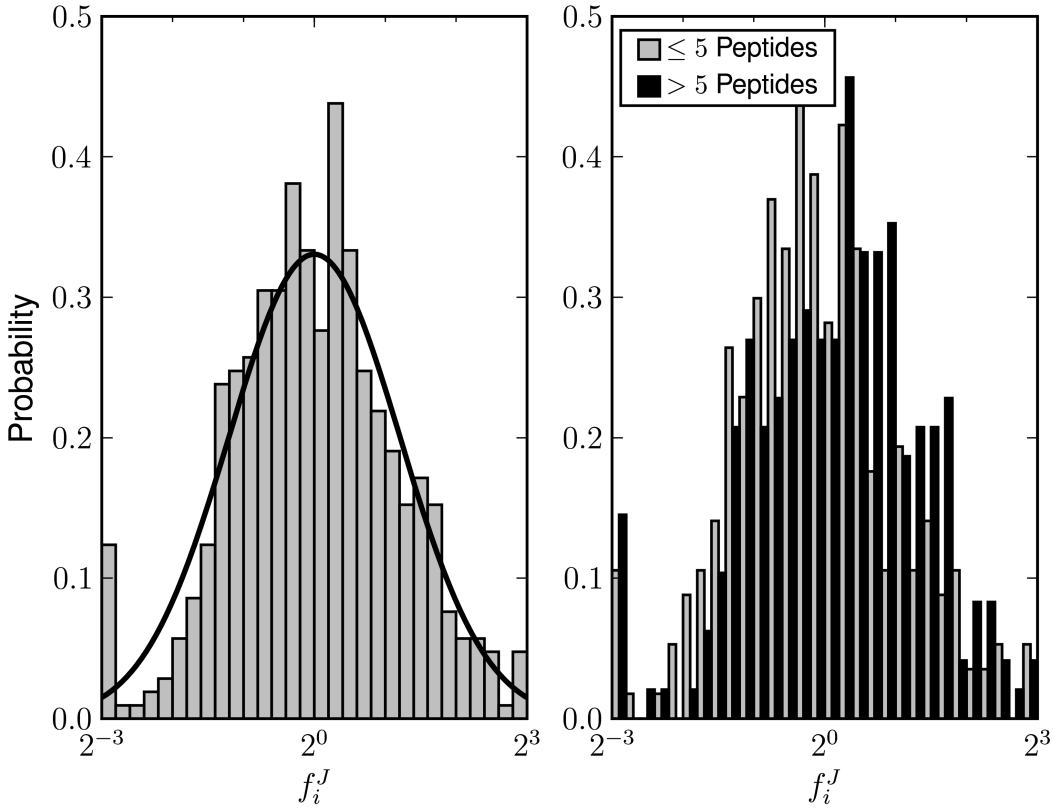

Figure 5.1: Scaled fold changes in specific protein synthesis expected from the experimentally observed mRNA and protein fold changes. Also shown in the left panel is a log normal distribution (line) fit to the results for visual reference only (the distribution fails a normality test). The right panel shows the same data, but split into more confident mean (number of peptides > 5) or less confidence mean (≤ 5 peptides) protein fold change. Both plots are scaled by the population geometric mean to center the distribution at 1.

mRNA ratios) or the biological regulation (undetermined distribution) was not able to be determined. Combined with the distribution failing a log normality test ($p = 2.2 \times 10^{-4}$), the unspecified distribution makes any comparison of the higher distribution modes statistically unsound. Thus, the results are presented visually as a histogram shown in the right panel of Figure 5.1. While the more confident proteins are shifted slightly right, both low and high confidence means appear to have nearly identical distribution widths,

with high confidence means perhaps showing a slightly greater width. The similarity of the high- and low-confidence mean distributions is consistent with the distribution width being a product of the underlying regulation of protein concentration (not experimental error), presenting a benchmark of two- to three-fold change in specific protein synthesis rates.

5.3.2 Predicted Changes in Synthesis Rates

The biochemical observations defining the ribosome and total mRNA concentration (see Methods) result in an average polysome size of fewer than ten ribosomes per mRNA. This polysome size is substantially below the theoretical maximum (approximately 65 per mRNA) or observed maximum with non rate-limiting translation initiation (approximately 15 per mRNA). The smaller average polysome size requires translation initiation be rate-limiting genome-wide to limit the expected number of actively translating 70S ribosomes. The change in translation initiation rates are shown in Figure 5.2, with a tight grouping and a negative skew representing those genes with significantly lower translation initiation efficiency. The predicted changes in protein synthesis rates, shown in Figure 5.3a, have a similar negative skew as a result of the rate-limiting translation-initiation and are significantly more narrowly distributed than experimentally observed in Figure 5.1.

Several hypothetical sets of translation kinetics are also explored. First, to explore the contribution of translation initiation, the results of setting translation initiation uniformly are shown in Figure 5.3b. The results lack the negative skew from changes in translation initiation (Figure 5.2). Next, two changes in the translation elongation rates were explored. First, to reflect the possible changes in translation elongation, the elon-

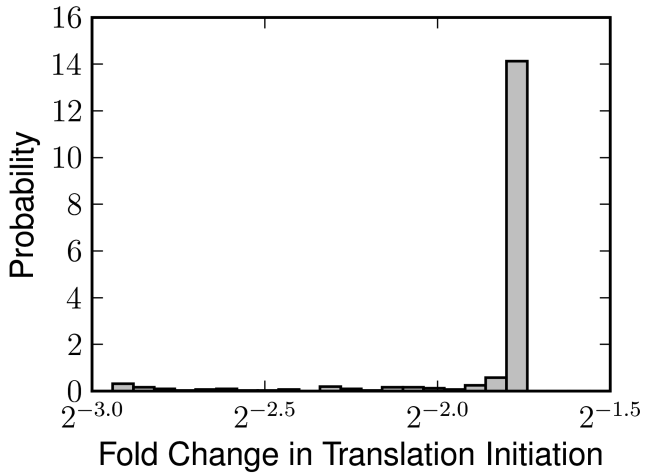

Figure 5.2: Changes in translation initiation rate expressed as the ratio of the latter (3.5 hours post-induction) to the initial translation initiation rate.

gation rates are randomly shuffled at the later data point, the results of which are shown in Figure 5.3c. These results show the widest distribution of changes in translation rates, but remain significantly more narrowly distributed than the experimentally observed. Second, to explore whether the distribution width of translation elongation rates affects the changes in protein synthesis rates, the spread of the translation elongation rate distribution was increased about the median rate (GAC) by either dividing lower rates or multiplying higher rates by 64. The results of this parameterization, shown in Figure 5.3d, are not significantly different than the original parameterization.

5.4 Discussion

Although no statistical significance can be assigned to the difference, the results shown in Figure 5.1 are consistent with the two- to three-fold spread in expected change in protein synthesis rates being a the result of the underlying translation regulation and not a product of experimental uncertainty. If true, this suggests that the model results shown

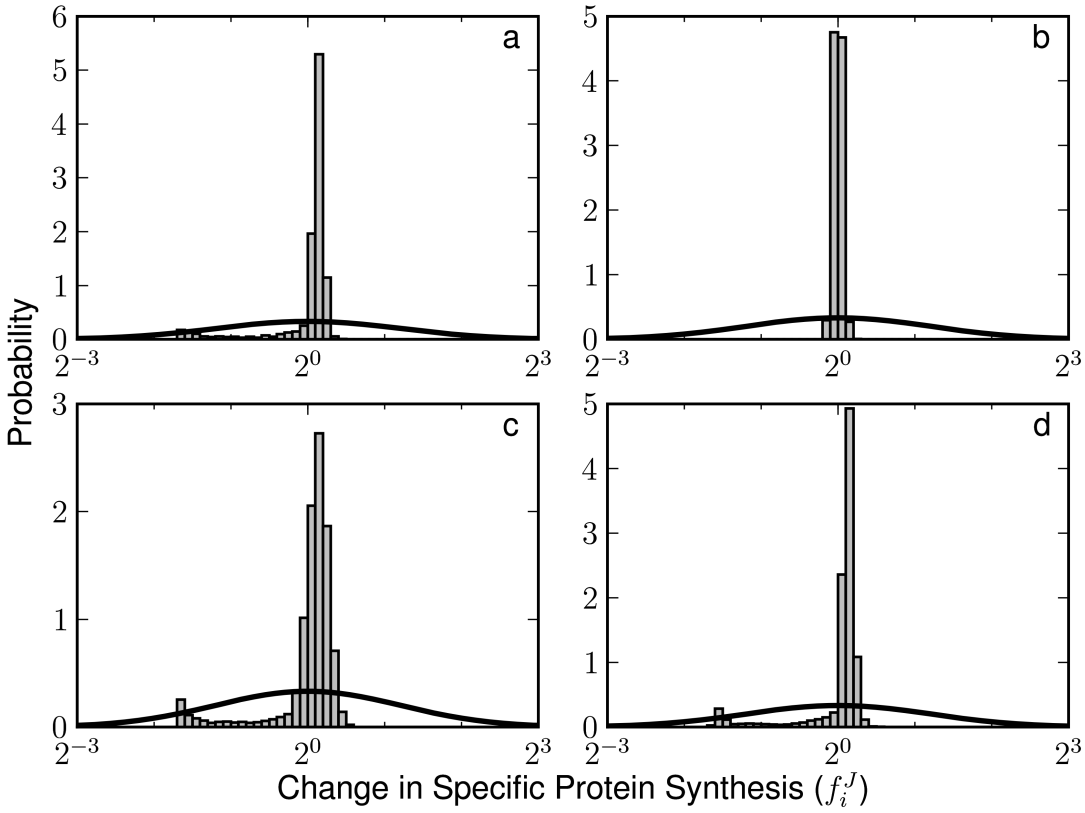

Figure 5.3: Scaled fold changes in specific protein synthesis predicted from the whole cell stochastic simulation (histogram) plotted against a log-normal distribution (curve) fit to the experimental observations using (a) the initiation and elongation kinetics defined in Methods, (b) uniform translation initiation, (c) randomly shuffled elongation rates at 3.5 hours, or (d) increased spread of the translation elongation rates about the median rate at 3.5 hours.

in Figure 5.3 for various possible parameterizations are too narrowly distributed. One possible regulatory mechanism which may increase this distribution is targeted protein degradation, implicitly assumed to be insignficant in the definition of Equation 5.1. If protein degradation is signficant relative to dilution by growth, the protein fold change of this protein would more quickly acheive steady state with respect to changes in mRNA. Unfortunately, little is known regarding the cell-wide targeting of proteins for degradation, and thus the implementation of targeted protein degradation may remain elu-

sive. The most likely parameterization further widening this distribution is expected to involve gene-specific levels of regulation and sequence specificity beyond codon bias such as better descriptions of changes in translation initiation. Thus, while global regulation of protein synthesis including translation elongation rates has garnered significant interest from both the experimental and computational communities, it appears more gene-specific knowledge of translational regulation is required to significantly improve our ability to predict the proteome composition. This gene-specific regulation may include binding of factors or riboswitches to modulate protein synthesis in response to cellular conditions.

5.5 Acknowledgments

We would like to thank Kunal Aggarwal for collecting the experimental observations and useful discussions regarding the resulting data normalization. We would also like to thank Sergios Nicolaou for useful discussions regarding the statistics of the experimentally observed changes in protein synthesis rates.

CHAPTER 6

EXACT DETERMINISTIC TRANSLATION MODEL

Cell-wide mathematical models of protein synthesis often determine the total cell-wide protein synthesis rate by balancing the predicted concentration of mRNA-bound actively translating ribosomes with the experimentally observed ribosome concentration. This calculation requires finding the average number of ribosomes bound to the mRNA, or polysome size, for each gene. Previous deterministic and mean-field models of protein synthesis were developed to predict the protein synthesis rates and encounter numerical issues at even moderate gene lengths preventing accurate polysome size predictions. An alternative model of protein synthesis, one based on mRNA states instead of ribosome states, is presented here. This model is an exact deterministic representation of all ribosome position combinations on the mRNA. Several simplifications are required to make the solution of moderately sized genes computationally obtainable, which can detract from the solution accuracy. Thus, while presenting a novel approach for calculating protein synthesis rates and polysome sizes with several potential benefits, the current model solution algorithm requires additional development to become a significantly better alternative to current mean-field solutions.

6.1 Introduction

Protein synthesis is a fundamental cellular process, producing much of the cell's functional content from enzymes to regulators. While measuring protein expression is important to the cellular state, the current proteome measurement techniques are expensive in time and resources and lack the genome-wide coverage of transcriptomic measurements. As a result, numerous mathematical models have been developed to describe the

complex process of protein synthesis [24, 80, 100]. These models incorporate diverse levels of complexity: from a detailed translation elongation description [100] to coarse-grained translation elongation [24, 80] to other translation regulators such as ribosome recycling [24]. These models are primarily applied to the prediction of the specific protein production rate of an individual gene, with relatively few works extending the model description to cell-wide protein synthesis. One cell-wide study of protein synthesis is that of Mehra and Hatzimanikatis [60]. To find the cell-wide protein synthesis rate, the authors balanced the cellular concentration of ribosomes with the predicted polysome size of each gene. The deterministic solution used has an error propagation issue [35, 79] which impacts the accuracy of polysome calculation for longer genes with higher polysome sizes. This numerical issue is problematic even in the relatively simple bacteria transcriptome where numerous genes have significantly inaccurate polysome size predictions. The most commonly used alternative is to calculate the polysome size and protein synthesis rate stochastically. Because stochastic solutions are costly in time, the full investigation of the translation kinetics is not possible. To address these problems, here we present an exact deterministic model of protein synthesis based on the mRNA state instead of the ribosome state. While the model description is relatively simple, the primary challenge is reducing the computation cost which grows exponentially with gene length. A method of splitting larger genes is presented along with a discussion of the remaining challenges and possible future efforts.

6.2 Model

The protein synthesis model presented here is unique in using an mRNA-based state description. In this description, each distinct combination of ribosome positions on an mRNA is considered as a single state. Because all distinct mRNA states are represented,

the solution of this model is an exact deterministic analog of the exact stochastic solution.

Protein synthesis is coarse-grained into three distinct steps: translation initiation, elongation, and termination. In these reactions, ribosomes are bound to the mRNA, move one codon (a nucleotide triplicate) to the 3' end of the mRNA, and dissociate from the mRNA after completing protein synthesis, respectively. Ribosomes are assumed to have an excluded volume of l codons, where $l = 12$ codons in the results presented below, which constrains two adjacent ribosomes to be separated by at minimum l codons. Each mRNA state may participate in one or more initiation, elongation, and terminations reactions to produce new distinct mRNA states. The resulting model description may be written as a series of differential equations. For example, the mRNA state with ribosomes at positions 15, 30, and 42 ($M_{\{15,30,42\}}$) is described by the following differential equation:

$$\frac{\mathrm{d}\, M_{\{15,30,42\}}}{\mathrm{dt}} = k_{14} \cdot M_{\{14,30,42\}} + k_{29} \cdot M_{\{15,29,42\}} - (k_{init} + k_{15} + k_{42}) \cdot M_{\{15,30,42\}} \qquad (6.1)$$

where k_{init} is the initiation rate and remaining k_* are the elongation rates at given mRNA positions. Note that $M_{\{15,30,42\}}$ cannot react to produce $M_{\{15,31,42\}}$ because this would violate the ribosome excluded volume between the second and third ribosome.

Assuming protein synthesis is at steady state, the model solution may be found by setting the linear system of differential equations to zero. The coefficient matrix, representing the reaction rate constants such as those shown in Equation 6.1, may be solved using linear algebraic tools. Because the coefficient matrix representation is rank $N - 1$, one arbitrarily chosen row constrains the sum of all mRNA states to 1. Using this constraint, the solution to the coefficient matrix gives the probability of finding each distinct mRNA state. The probability of finding ribosome at position i, also called a ribosome density (ρ_i), is defined as the cumulative probability of all distinct mRNA states with a

ribosome at position i. The summation of ribosome densities ρ_i along the entire mRNA defines the polysome size. Additionally, the protein synthesis rate may be found from the product of the termination rate at the last ribosome density ($J = k_{term} \cdot \rho_L$).

As a result of the computational requirements of the even moderate gene lengths (see Computational Requirements below), the typical bacterium gene cannot be directly solved. To overcome this limitation, large gene sequences may be divided into a set of smaller subsequences. Because the subsequence is part of a larger sequence, the pseudo-initiation or -termination reactions for one subsequence depend upon all the other subsequences. The challenge then becomes how to properly account for these edge-effects in split sequences. To exactly solve for the edge-effects, the larger coefficient matrix could be reduced so mRNA states within a given subsequence only react to and are produced from other mRNA states in the same subsequence. As expected, this solution is similar in computational effort to solving the entire message length and thus not a tractable solution.

A simple treatment of the edge-effects assumes each subsequence is individually at steady state (i.e. ribosome movement through the surrounding frames is infinitely fast). Effectively, this assumption allows each subsequence to be treated as an independent mRNA. The solution to the entire mRNA may then be reassembled from the solution of the independent pieces. The algorithm used to solve the model is presented below.

1. Split the gene into subsequences of 60 codons with overlap of $2 \cdot l = 24$ codons between neighboring subsequences.
2. Solve each subsequence, finding the slowest specific protein production rate, also known as the current.
3. Iteratively solve all other subsequences with faster currents by bisecting the initiation rate constant (for subsequences 3' of the slowest subsequence) or the ter-

mination rate constant (for subsequences 5' of the slowest subsequence) until the calculated current converges to the slowest.

6.3 Results

6.3.1 Computational Requirements

The number of unique ribosome position combinations on a message where the furthest 3' ribosome is at position i can be expressed by a linear recursion equation.

$$n(i) = \begin{cases} 1 & 1 \leq i \leq l \\ n(i-1) + n(i-l) & i > l \end{cases} \tag{6.2}$$

where l is the excluded volume (i.e. size) of the ribosome. From the results of this equation, the total number of mRNA states required (N) for an mRNA of L codons may be found:

$$N(L) = 1 + \sum_{i=1}^{L} n(i) \tag{6.3}$$

The positions count are indexed from 1 at the 5' end of the mRNA to L at the 3' end, with an additional state included to represent the empty mRNA with no bound ribosomes.

The steady state solution of the linear system of $N(L)$ states is expressed as the solution to the coefficient matrix. The number of non-zero entries in the coefficient matrix may be calculated using the following linear recursion equation.

$$M(L) = \begin{cases} 3 \cdot N(L) - 2 & 0 \leq L \leq l \\ M(L-1) + M(L-l) + n(L) + n(L-1) - N(L-l) + 2 & L > 1 \end{cases} \tag{6.4}$$

As seen in Figure 6.1 for $l = 12$, the memory requirement of the coefficient matrix becomes intractable at a message length of 50–100 codons.

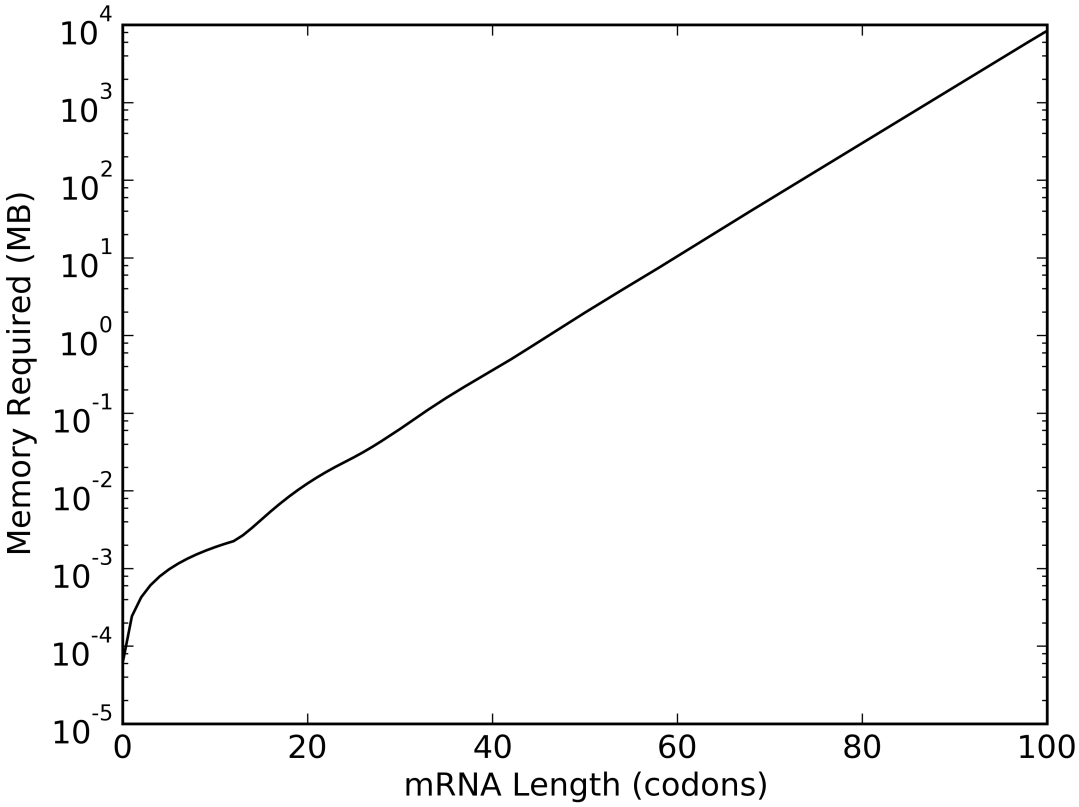

Figure 6.1: Lowest possible memory requirement for storing a sparse coefficient matrix for the model solution. The memory requirement assumes a 64-bit double precision for each non-zero entry in the coefficient matrix ($M(L)$).

6.3.2 Comparison to Mean-Field Model

The results for the mRNA-based model presented here are compared to the protein synthesis rate (current) and polysome size calculated from the mean-field solution [80] for a random set of *Escherichia coli* genes from [80]. The translation initiation and elongation rates were set non-limiting and the elongation rates were taken from Fluitt et al. [21].

As shown in Table 6.1, both the mean-field and the mRNA-based model predictions

provide reasonably accurate predictions of the protein synthesis rates. The mRNA-based model predicts consistently higher protein synthesis rates as a result of splitting the sequence and thus not capturing all rate-limiting kinetics in a single subsequence. Unfortunately, both models fail to consistently and accurately predict the polysome size, although the mRNA-based model appears to perform moderately better.

Figure 6.2 shows the ribosome coverage of the longest gene, *aceF*, where ribosome coverage is defined as the cumulative ribosome density over l codons. While the mean-field solution does not initially appear to encounter numerical issues, the sudden divergence of the ribosome coverage profile from the stochastic solutions illustrates an intermediate case where the mean-field solution quietly fails. Conversely, the mRNA-based model predicts ribosome coverage more accurately where the lack of precision not perpetuated through the calculation. The unphysical ribosome coverage greater than one, shown for the mRNA-based model results, is a result of joining the edges of independent subsequences.

6.4 Discussion and Conclusions

Both the mean-field solution and the mRNA-based solution provide reasonable accurate predictions of the protein synthesis rate. The latter result suggests the relatively small subsequence size used here (60 codons) was able to capture a majority of the rate-limiting translation kinetics. If the mRNA-based model is used solely for the prediction of protein synthesis rates, an algorithm iteratively finding the slowest subsequence position may produce more accurate results.

As expected, the mean-field solution has significant inaccuracies in the predicted polysome size. Much of this error results from unknown ribosome density to the 5'

Table 6.1: Errors in the predicted protein synthesis rate and polysome size for randomly selected *E. coli* genes.

Gene	Length	Current	% Error	
		Polysome	Mean-Field	mRNA-Based
crr	170	27.11	-1.34	2.48
		10.15	-5.32	-6.72
adk	215	28.10	-1.03	1.29
		11.86	2.67	-11.15
fabG	245	25.30	5.35	5.42
		14.75	-3.84	-5.00
glnH	249	23.12	-1.11	0.25
		15.80	16.97†	-1.50
envY	254	17.77	-3.13	0.94
		14.04	7.11†	2.52
fabD	310	23.41	-1.37	3.10
		17.17	2.07†	-1.55
cysK	324	23.58	-0.06	2.44
		19.57	-17.20	-2.22
gapA	332	26.76	2.52	3.98
		20.10	8.19†	1.28
asnS	467	23.10	2.04	2.64
		25.43	9.06†	-5.40
aceF	631	24.28	-2.73	2.52
		39.86	-31.00	5.30

Notes: The current and polysome size presented are from a stochastic solution run for 10^7 proteins produced. % Error is defined as $100 \cdot (\textit{prediction} - \textit{actual})/\textit{actual}$. † denotes mean-field solutions with numerical error propagation issues. The ribosomes densities 5' of the mean-field error were set to $1/l$.

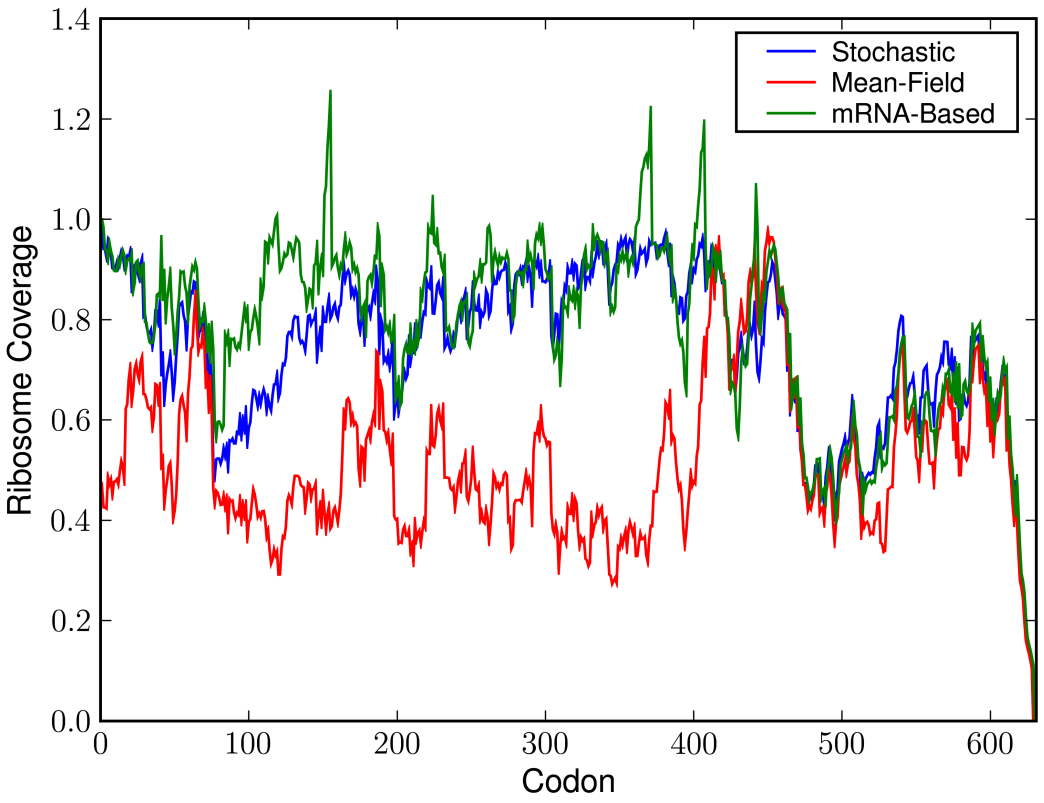

Figure 6.2: Ribosome coverage for the typical *E. coli* gene *aceF* from the mean-field and mRNA-based models, defined as the cumulative ribosome density over $l = 12$ codons.

end of the mRNA after encountering a round-off error. Unfortunately, the approximation presented for the mRNA-based model provides only moderately more accurate polysome size predictions. In contrast to the mean-field solution, these inaccuracies are not a result of numerical precision of the calculation but instead the approximation used when splitting a large gene sequence. Thus, the assumption that each subsequence is at steady state may not yield sufficiently accurate predictions.

Several future developments may be used to improve the prediction accuracy of this model. One future development is the use of a heuristic approach whereby a series of slow rates a fixed distance outside the subsequence are studied and used to estimate the

edge-effects for real gene sequences. This approach could avoid the numerical challenge of solving the entire gene while providing a possible improvement in the polysome size prediction. Another future development is the selective discarding of unimportant translation elongation steps, which essentially would eliminate the mRNA states which have a probability significantly less than the most probable mRNA states. The challenge with this approach will be consistently reducing the number of model states to a computationally tractable size.

Here we present a novel exact deterministic representation of protein synthesis. This model representation allows the exact calculation of protein synthesis rates and ribosome density profile, avoiding the numerical issues of the mean-field deterministic solution. Unfortunately, the numerical precision challenge of the mean-field solution is substituted for by the computational challenge of solving the large number of distinct mRNA states. This challenge may be addressed by coarse-graining larger sequences into smaller solvable subsequences, which we accomplish here with an overlapping set of subsequences. Because exactly solving the edge-effects (pseudo-initiation and -elongation) for these subsequences is as computationally expensive as the original solution, a simplistic approach ignoring the rate-limiting effects outside each subsequence was presented. The results show this approximate model solution is moderately more accurate than the mean-field solution, but future development is necessary to provide a consistently better predictive capability.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

We have presented two computational biology efforts to elicit greater understanding of the regulation circadian rhythmicity and the prediction of the bacterial proteome. We summarize the results of both projects and suggest some possible future investigations in this chapter.

7.1 Summary

In the first effort we addressed the open questions of circadian regulation and communication which have been raised in the literature. We find that independent nuclear transfer of central negative feedback of PERIOD and TIMELESS may require an additional feedback loop to produce the switch-like dissociation observed experimentally. This positive feedback was introduced into a mechanistic mathematical framework for *Drosophila* circadian rhythms which demonstrates excellent agreement with experimentally observed expression profiles of circadian genes and many circadian mutants. The understanding gained from the single pacemaker was then extended to study the pacemaker communication underlying seasonal adaptation of *Drosophila* locomotor activity. Using a systematic search for the currently unknown communication mechanisms, the model results illustrate a possible coupling mechanism including a novel circadian signal FDP complementing the well-studied PDF neuropeptide. The proposed mechanism presents novel insights into pacemaker communication which may lead to future experimentation and ultimately advancements in regulating human circadian communication.

The second effort extends our previously developed model of protein synthesis to genome-wide measurements. From relatively inexpensive microarray measurements of

transcriptomic changes, this model can predict the expected proteome using heterogeneous translation initiation and elongation kinetics reflecting gene-specific differences in the ribosome binding site and codon bias. The model predictions are consistent with the experimentally observed proteomic response to *rhsA* over-expression in *Escherichia coli*. Genes which are not well described by this model may be post-translationally regulated by, among other possibilities, riboswitches and targeted protein degradation. The experimental observations are also shown to have a significantly wider distribution of changes in specific protein synthesis rates than model results. This result re-emphasizes the need to incorporate better gene-specific regulation to improve the predicted proteome. As a final part of this effort, a new model of protein synthesis is presented which represents the exact deterministic representation of translation. While presenting greater potential for computationally inexpensive solutions with improved accuracy than the mean-field solution, further development of this model is required.

7.2 Future Work

7.2.1 Circadian Rhythmicity

Circadian rhythmicity continues to be a hot research topic because of the fundamental interest in biological oscillations. As a result, numerous (mostly academic) experimental and computational efforts generate a large body of work each year. Without direct collaboration with experimental efforts, any computational biology effort thrives almost exclusively on published results. These published results may have been first observed six months to one year prior to the experimental publication. With a modest turnaround time of one year to have the relevant computational analysis and hypotheses published,

the circular paradigm of systems biology will take approximately two years per oscillation. Given the fast pace of circadian rhythms research and this long turn-around time, the closer collaboration of any computational effort with both the computational and experimental circadian communities will greatly improve the project's impact. Collaboration with computational groups will aid in preventing redundant investigations, whereas experimental collaboration will increase the impact of computational results while possibly yielding access to unpublished (often interesting) observations. Thus, the future impact of this project may rely upon opening a dialogue to potential collaborators.

The communication of pacemaker cells should remain at the forefront of circadian research because of its direct effect on observed circadian physiology (e.g. seasonal adaption to day length) and health-related effects (e.g. jet lag). Thus a redefined search for possible communication signals and the resulting actions may provide better insight into the interaction between pacemaker cells. The proposed connectivity resulting from the genetic algorithm search in Chapter 3 fails to show morning and evening oscillator codominance with varying day length. There are several possible explanations for this shortcoming. First, the partitioning of evening and morning oscillators may prevent the accurate search of reaction space. A search using the whole interacting network may provide a more accurate search, but may produce unstable children more frequently, increasing the computational cost. Second, the deterministic solution may prevent the optimization algorithm from exploring unstable morning or evening oscillators. Using a stochastic solution to the model, while at a greater computational expense to find the average locomotor activity, may facilitate the inclusion of unstable pacemakers. Finally, the relative differences of the locomotor activity profiles may be too subtle to capture using amplitude data alone. Additional χ residuals directly constraining the relative shifts of the locomotor activity peaks may facilitate finding a codominant communication mechanism. These residuals could also reflect the sensitivity of the morning peak

in locomotor activity to changes in the evening oscillator pace and vice versa. These alternative optimization algorithms may yield better insight into pacemaker codominance under varying day lengths.

7.2.2 Protein Synthesis

As presented in Chapter 5, the observed distribution of experimentally observed changes in specific protein synthesis rate is consistent with underlying post-transcriptional regulation. Relative to these experimental observations, the model results are significantly more narrowly distributed. Thus types of translation regulation beyond the current inhomogeneous translation initiation and elongation kinetics present an opportunity to increase the distribution width of protein synthesis rates and possibly the accuracy of model predictions. One particularly interesting possibility is riboswitch regulation wherein metabolites and other small molecules can regulate the translational activity of a transcript. The incorporation of this level of regulation would most likely involve detailing the cellular state of metabolite and small molecules regulators to describe translational regulation under changing cellular conditions. This additional complexity may improve the accuracy of the predicted proteome and provide an insight into the relative significance of different forms of translation regulation.

BIBLIOGRAPHY

[1] K Aggarwal, L H Choe, and K H Lee. Quantitative analysis of protein expression using amine-specific isobaric tags in *Escherichia coli* cells expressing *rhsA* elements. *Proteomics*, 5(9):2297–2308, Jun 2005.

[2] K Aggarwal and K H Lee. Over-expression of *rhsA* elements perturbs the cellular translational machinery in *Escherichia coli*. *Paper in Preparation*, 2008.

[3] R Allada, N E White, W V So, J C Hall, and M Rosbash. A mutant *Drosophila* homolog of mammalian *Clock* disrupts circadian rhythms and transcription of *period* and *timeless*. *Cell*, 93(5):791–804, May 1998.

[4] L Anderson and J Seilhamer. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, 18(3–4):533–537, Mar–Apr 1997.

[5] L J Ashmore, S Sathyanarayanan, D W Silvestre, M M Emerson, P Schotland, and A Sehgal. Novel insights into the regulation of the Timeless protein. *J Neurosci*, 23(21):7810–7819, Aug 2003.

[6] K Bae, C Lee, P E Hardin, and I Edery. dCLOCK is present in limiting amounts and likely mediates daily interactions between the dCLOCK-CYC transcription factor and the PER-TIM complex. *J Neurosci*, 20(5):1746–1753, Mar 2000.

[7] K Bae, C Lee, D Sidote, K Y Chuang, and I Edery. Circadian regulation of a *Drosophila* homolog of the mammalian *Clock* gene: PER and TIM function as positive regulators. *Mol Cell Biol*, 18(10):6142–6151, Oct 1998.

[8] M K Baylies, T A Bargiello, F R Jackson, and M W Young. Changes in abundance or structure of the *per* gene product can alter periodicity of the *Drosophila* clock. *Nature*, 326(6111):390–392, Mar 1987.

[9] J Blau and M W Young. Cycling *vrille* expression is required for a functional *Drosophila* clock. *Cell*, 99(6):661–671, Dec 1999.

[10] Y Bobinnec, X Morin, and A Debec. Shaggy/GSK-3β kinase localizes to the centrosome and to specialized cytoskeletal structures in *Drosophila*. *Cell Motil Cytoskeleton*, 63(6):313–320, Jun 2006.

[11] H Bolouri and E H Davidson. Transcriptional regulatory cascades in development: Initial rates, not steady state, determine network kinetics. *Proc Natl Acad Sci U S A*, 100(16):9371–9376, Aug 2003.

[12] J Bongard and H Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proc Natl Acad Sci U S A*, 104(24):9943–9948, Jun 2007.

[13] H Bremer and P P Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. In Frederick C. Neidhardt, editor, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, volume 2, chapter 97, pages 1553–1569. Washington, D.C., 2nd edition, 1996.

[14] D C Chang and S M Reppert. A novel C-terminal domain of *Drosophila* PERIOD inhibits dCLOCK:CYCLE-mediated transcription. *Curr Biol*, 13(9):758–762, Apr 2003.

[15] P Cheng, Y Yang, and Y Liu. Interlocked feedback loops contribute to the robustness of the *Neurospora* circadian clock. *Proc Natl Acad Sci U S A*, 98(13):7408–7413, Jun 2001.

[16] B H Collins, E Rosato, and C P Kyriacou. Seasonal behavior in *Drosophila melanogaster* requires the photoreceptors, the circadian clock, and phospholipase C. *Proc Natl Acad Sci U S A*, 101(7):1945–1950, Feb 2004.

[17] S A Cyran, A M Buchsbaum, K L Reddy, M C Lin, N R Glossop, P E Hardin, M W Young, R V Storti, and J Blau. *vrille*, *Pdp1*, and *dClock* form a second feedback loop in the *Drosophila* circadian clock. *Cell*, 112(3):329–341, Feb 2003.

[18] S A Cyran, G Yiannoulos, A M Buchsbaum, L Saez, M W Young, and J Blau. The Double-Time protein kinase regulates the subcellular localization of the *Drosophila* clock protein Period. *J Neurosci*, 25(22):5430–5437, Jun 2005.

[19] M H de Smit and J van Duin. Translational standby sites: How ribosomes may deal with the rapid folding kinetics of mRNA. *J Mol Biol*, 331(4):737–743, Aug 2003.

[20] I Edery, L J Zwiebel, M E Dembinska, and M Rosbash. Temporal phosphorylation of the *Drosophila* Period protein. *Proc Natl Acad Sci U S A*, 91(6):2260–2264, Mar 1994.

[21] A Fluitt, E Pienaar, and H Viljoen. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput Biol Chem*, 31(5–6):335–346, Oct 2007.

[22] L Fu and C C Lee. The circadian clock: Pacemaker and tumour suppressor. *Nat Rev Cancer*, 3(5):350–361, May 2003.

[23] A Gattiker, E Gasteiger, and A Bairoch. ScanProsite: A reference implementation of a PROSITE scanning tool. *Appl Bioinformatics*, 1(2):107–108, 2002.

[24] M A Gilchrist and A Wagner. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *J Theor Biol*, 239(4):417–434, Apr 2006.

[25] N R Glossop, J H Houl, H Zheng, F S Ng, S M Dudek, and P E Hardin. VRILLE feeds back to control circadian transcription of *Clock* in the *Drosophila* circadian oscillator. *Neuron*, 37(2):249–261, Jan 2003.

[26] N R Glossop, L C Lyons, and P E Hardin. Interlocked feedback loops within the *Drosophila* circadian oscillator. *Science*, 286(5440):766–768, Oct 1999.

[27] D Gonze, J Halloy, and A Goldbeter. Robustness of circadian rhythms with respect to molecular noise. *Proc Natl Acad Sci U S A*, 99(2):673–678, Jan 2002.

[28] B Grima, E Chélot, R Xia, and F Rouyer. Morning and evening peaks of activity rely on different clock neurons of the *Drosophila* brain. *Nature*, 431(7010):869–873, Oct 2004.

[29] R N Gutenkunst, F P Casey, J J Waterfall, J C Atlas, R S Kuczenski, C R Myers, and J P Sethna. SloppyCell. `http://sloppycell.sourceforge.net`.

[30] R N Gutenkunst, J J Waterfall, F P Casey, K S Brown, C R Myers, and J P Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10):1871–1878, Oct 2007.

[31] S P Gygi, Y Rochon, B R Franza, and R Aebersold. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, 19(3):1720–1730, Mar 1999.

[32] P E Hardin, J C Hall, and M Rosbash. Circadian oscillations in *period* gene mRNA levels are transcriptionally regulated. *Proc Natl Acad Sci U S A*, 89(24):11711–11715, Dec 1992.

[33] M H Hastings. Circadian clockwork: Two loops are better than one. *Nat Rev Neurosci*, 1(2):143–146, Nov 2000.

[34] M H Hastings, A B Reddy, and E S Maywood. A clockwork web: Circadian timing in brain and periphery, in health and disease. *Nat Rev Neurosci*, 4(8):649–661, Aug 2003.

[35] R Heinrich and T A Rapoport. Mathematical modelling of translation of mRNA in eucaryotes; steady state, time-dependent processes and application to reticulocytes. *J Theor Biol*, 86(2):279–313, Sep 1980.

[36] M Hunter-Ensor, A Ousley, and A Sehgal. Regulation of the *Drosophila* protein Timeless suggests a mechanism for resetting the circadian clock by light. *Cell*, 84(5):677–685, Mar 1996.

[37] S Hyun, Y Lee, S T Hong, S Bang, D Paik, J Kang, J Shin, J Lee, K Jeon, S Hwang, E Bae, and J Kim. *Drosophila* GPCR Han is a receptor for the circadian clock neuropeptide PDF. *Neuron*, 48(2):267–278, Oct 2005.

[38] T Ideker, V Thorsson, J A Ranish, R Christmas, J Buhler, J K Eng, R Bumgarner, D R Goodlett, R Aebersold, and L Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934, May 2001.

[39] S Kadener, D Stoleru, M McDonald, P Nawathean, and M Rosbash. *Clockwork Orange* is a transcriptional repressor and a new *Drosophila* circadian pacemaker component. *Genes Dev*, 21(13):1675–1686, Jul 2007.

[40] E Y Kim, K Bae, F S Ng, N R Glossop, P E Hardin, and I Edery. *Drosophila* CLOCK protein is under posttranscriptional control and influences light-induced activity. *Neuron*, 34(1):69–81, Mar 2002.

[41] E Y Kim and I Edery. Balance between DBT/CKIϵ kinase and protein phosphatase activities regulate phosphorylation and stability of *Drosophila* CLOCK protein. *Proc Natl Acad Sci U S A*, 103(16):6178–6183, Apr 2006.

[42] R J Konopka and S Benzer. Clock mutants of *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*, 68(9):2112–2116, Sep 1971.

[43] R S Kuczenski, K C Hong, J García-Ojalvo, and K H Lee. Mathematical model of *Drosophila* circadian rhythmicity. `http://www.leelab.org/mathematical_models/RobertKuczenski/CircadianRhythmsModel`.

[44] R S Kuczenski, K C Hong, J García-Ojalvo, and K H Lee. PERIOD-TIMELESS interval timer may require an additional feedback loop. *PLoS Comput Biol*, 3(8):e154, Aug 2007.

[45] R S Kuczenski and K H Lee. Mathematical model of the seasonal adaptation of

Drosophila circadian rhythmicity. http://www.leelab.org/mathematical_models/RobertKuczenski/SeasonalAdaptation.

[46] B C Lear, C E Merrill, J M Lin, A Schroeder, L Zhang, and R Allada. A G protein-coupled receptor, *groom-of-PDF*, is required for PDF neuron action in circadian behavior. *Neuron*, 48(2):221–227, Oct 2005.

[47] C Lee, K Bae, and I Edery. The *Drosophila* CLOCK protein undergoes daily rhythms in abundance, phosphorylation, and interactions with the PER-TIM complex. *Neuron*, 21(4):857–867, Oct 1998.

[48] P S Lee and K H Lee. Engineering HlyA hypersecretion in *Escherichia coli* based on proteomic and microarray analyses. *Biotechnol Bioeng*, 89(2):195–205, Jan 2005.

[49] P S Lee, L B Shaw, L H Choe, A Mehra, V Hatzimanikatis, and K H Lee. Insights into the relation between mRNA and protein expression patterns: II. Experimental observations in *Escherichia coli*. *Biotechnol Bioeng*, 84(7):834–841, Dec 2003.

[50] J C Leloup and A Goldbeter. A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *J Biol Rhythms*, 13(1):70–87, Feb 1998.

[51] C Lim, B Y Chung, J L Pitman, J J McGill, S Pradhan, J Lee, K P Keegan, J Choe, and R Allada. *clockwork orange* encodes a transcriptional repressor important for circadian-clock amplitude in *Drosophila*. *Curr Biol*, 17(12):1082–1089, Jun 2007.

[52] Y Lin, G D Stormo, and P H Taghert. The neuropeptide Pigment-Dispersing Factor coordinates pacemaker interactions in the *Drosophila* circadian system. *J Neurosci*, 24(36):7951–7957, Sep 2004.

[53] J Majercak, D Sidote, P E Hardin, and I Edery. How a circadian clock adapts to seasonal decreases in temperature and day length. *Neuron*, 24(1):219–230, Sep 1999.

[54] M Mandal, B Boese, J E Barrick, W C Winkler, and R R Breaker. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell*, 113(5):577–586, May 2003.

[55] M Mandal, M Lee, J E Barrick, Z Weinberg, G M Emilsson, W L Ruzzo, and R R

Breaker. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science*, 306(5694):275–279, Oct 2004.

[56] N R Markham and M Zuker. UNAFold: Software for nucleic acid folding and hybriziation. In J M Keith, editor, *Bioinformatics, Volume II. Structure, Functions and Applications (Methods in Molecular Biology)*, volume 453, chapter 1, pages 3–31. 2008.

[57] S B Marrus, H Zeng, and M Rosbash. Effect of constant light and circadian entrainment of per^S flies: Evidence for light-mediated delay of the negative feedback loop in *Drosophila*. *EMBO J*, 15(24):6877–6886, Dec 1996.

[58] S Martinek, S Inonog, A S Manoukian, and M W Young. A role for the segment polarity gene *shaggy*/GSK-3 in the *Drosophila* circadian clock. *Cell*, 105(6):769–779, Jun 2001.

[59] A Matsumoto, M Ukai-Tadenuma, R G Yamada, J Houl, K D Uno, T Kasukawa, B Dauwalder, T Q Itoh, K Takahashi, R Ueda, P E Hardin, T Tanimura, and H R Ueda. A functional genomics strategy reveals *clockwork orange* as a transcriptional regulator in the *Drosophila* circadian clock. *Genes Dev*, 21(13):1687–1700, Jul 2007.

[60] A Mehra and V Hatzimanikatis. An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophys J*, 90(4):1136–1146, Feb 2006.

[61] I Mertens, A Vandingenen, E C Johnson, O T Shafer, W Li, J S Trigg, A De Loof, L Schoofs, and P H Taghert. PDF receptor signaling in *Drosophila* contributes to both circadian and geotactic behaviors. *Neuron*, 48(2):213–219, Oct 2005.

[62] P Meyer, L Saez, and M W Young. PER-TIM interactions in living *Drosophila* cells: An interval timer for the circadian clock. *Science*, 311(5758):226–229, Jan 2006.

[63] A Murad, M Emery-Le, and P Emery. A subset of dorsal neurons modulates circadian behavior and light responses in *Drosophila*. *Neuron*, 53(5):689–701, Mar 2007.

[64] M P Myers, K Wager-Smith, A Rothenfluh-Hilfiker, and M W Young. Light-induced degradation of TIMELESS and entrainment of the *Drosophila* circadian clock. *Science*, 271(5256):1736–1740, Mar 1996.

[65] P Nawathean and M Rosbash. The Doubletime and CKII kinases collaborate to potentiate *Drosophila* PER transcriptional repressor activity. *Mol Cell*, 13(2):213–223, Jan 2004.

[66] J H Park, C Helfrich-Förster, G Lee, L Liu, M Rosbash, and J C Hall. Differential regulation of circadian pacemaker output by separate clock genes in *Drosophila*. *Proc Natl Acad Sci U S A*, 97(7):3608–3613, Mar 2000.

[67] C S Pittendrigh and S Daan. A functional analysis of circadian pacemakers in nocturnal rodents V. Pacemaker structure: A clock for all seasons. *J Comp Physiol A*, 106(3):333–355, Oct 1976.

[68] J L Price, J Blau, A Rothenfluh, M Abodeely, B Kloss, and M W Young. *double-time* is a novel *Drosophila* clock gene that regulates PERIOD protein accumulation. *Cell*, 94(1):83–95, Jul 1998.

[69] S C Renn, J H Park, M Rosbash, J C Hall, and P H Taghert. A *pdf* neuropeptide gene mutation and ablation of PDF neurons each cause severe abnormalities of behavioral circadian rhythms in *Drosophila*. *Cell*, 99(7):791–802, Dec 1999.

[70] D Rieger, O T Shafer, K Tomioka, and C Helfrich-Förster. Functional analysis of circadian pacemaker neurons in *Drosophila melanogaster*. *J Neurosci*, 26(9):2531–2543, Mar 2006.

[71] P L Ross, Y N Huang, J N Marchese, B Williamson, K Parker, S Hattan, N Khainovski, S Pillai, S Dey, S Daniels, S Purkayastha, P Juhasz, S Martin, M Bartlet-Jones, F He, A Jacobson, and D J Pappin. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*, 3(12):1154–1169, Dec 2004.

[72] A Rothenfluh, M Abodeely, J L Price, and M W Young. Isolation and analysis of six *timeless* alleles that cause short- or long-period circadian rhythms in *Drosophila*. *Genetics*, 156(2):665–675, Oct 2000.

[73] A Rothenfluh, M Abodeely, and M W Young. Short-period mutations of *per* affect a *double-time*-dependent step in the *Drosophila* circadian clock. *Curr Biol*, 10(21):1399–1402, Nov 2000.

[74] A Rothenfluh, M W Young, and L Saez. A TIMELESS-independent function for PERIOD proteins in the *Drosophila* clock. *Neuron*, 26(2):505–514, May 2000.

[75] J E Rutila, V Suri, M Le, W V So, M Rosbash, and J C Hall. CYCLE is a second

bHLH-PAS clock protein essential for circadian rhythmicity and transcription of *Drosophila period* and *timeless*. *Cell*, 93(5):805–814, May 1998.

[76] T Scheper, D Klinkenberg, C Pennartz, and J van Pelt. A mathematical model for the intracellular circadian rhythm generator. *J Neurosci*, 19(1):40–47, Jan 1999.

[77] A Sehgal, A Rothenfluh-Hilfiker, M Hunter-Ensor, Y Chen, M P Myers, and M W Young. Rhythmic expression of *timeless*: A basis for promoting circadian cycles in *period* gene autoregulation. *Science*, 270(5237):808–810, Nov 1995.

[78] O T Shafer, M Rosbash, and J W Truman. Sequential nuclear accumulation of the clock proteins Period and Timeless in the pacemaker neurons of *Drosophila melanogaster*. *J Neurosci*, 22(14):5946–5954, Jul 2002.

[79] L B Shaw, J P Sethna, and K H Lee. Mean-field approaches to the totally asymmetric exclusion process with quenched disorder and large particles. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(2 Pt 1):021901, Aug 2004.

[80] L B Shaw, R K Zia, and K H Lee. Totally asymmetric exclusion process with extended objects: A model for protein synthesis. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(2 Pt 1):021910, Aug 2003.

[81] R F Smith and R J Konopka. Effects of dosage alterations at the *per* locus on the period of the circadian clock of *Drosophila*. *Mol Gen Genet*, 185(1):30–36, Mar 1982.

[82] P Smolen, D A Baxter, and J H Byrne. Modeling circadian oscillations with interlocking positive and negative feedback loops. *J Neurosci*, 21(17):6644–6656, Sep 2001.

[83] P Smolen, P E Hardin, B S Lo, D A Baxter, and J H Byrne. Simulation of *Drosophila* circadian oscillations, mutations, and light responses by a model with VRI, PDP-1, and CLK. *Biophys J*, 86(5):2786–2802, May 2004.

[84] W V So and M Rosbash. Post-transcriptional regulation contributes to *Drosophila* clock gene mRNA cycling. *EMBO J*, 16(23):7146–7155, Dec 1997.

[85] S V Spinelli, L B Pontel, E García Véscovi, and F C Soncini. Regulation of magnesium homeostasis in *Salmonella*: Mg(2+) targets the *mgtA* transcript for degradation by RNase E. *FEMS Microbiol Lett*, 280(2):226–234, Mar 2008.

[86] R Stanewsky. Genetic analysis of the circadian system in *Drosophila melanogaster* and mammals. *J Neurobiol*, 54(1):111–147, Jan 2003.

[87] R Stanewsky, K S Lynch, C Brandes, and J C Hall. Mapping of elements involved in regulating normal temporal *period* and *timeless* RNA expression patterns in *Drosophila melanogaster*. *J Biol Rhythms*, 17(4):293–306, Aug 2002.

[88] J Stelling, E D Gilles, and F J Doyle. Robustness properties of circadian clock architectures. *Proc Natl Acad Sci U S A*, 101(36):13210–13215, Sep 2004.

[89] D Stoleru, P Nawathean, M P Fernández, J S Menet, M F Ceriani, and M Rosbash. The *Drosophila* circadian network is a seasonal timer. *Cell*, 129(1):207–219, Apr 2007.

[90] B L Taylor and I B Zhulin. PAS domains: Internal sensors of oxygen, redox potential, and light. *Microbiol Mol Biol Rev*, 63(2):479–506, Jun 1999.

[91] J J Tyson, C I Hong, C D Thron, and B Novak. A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *Biophys J*, 77(5):2411–2417, Nov 1999.

[92] H R Ueda, M Hagiwara, and H Kitano. Robust oscillations within the interlocked feedback model of *Drosophila* circadian rhythm. *J Theor Biol*, 210(4):401–406, Jun 2001.

[93] H R Ueda, K Hirose, and M Iino. Intercellular coupling mechanism for synchronized and noise-resistant circadian oscillators. *J Theor Biol*, 216(4):501–512, Jun 2002.

[94] L B Vosshall, J L Price, A Sehgal, L Saez, and M W Young. Block in nuclear localization of *period* protein by a second clock mutation, *timeless*. *Science*, 263(5153):1606–1609, Mar 1994.

[95] J Waterhouse, T Reilly, G Atkinson, and B Edwards. Jet lag: Trends and coping strategies. *Lancet*, 369(9567):1117–1129, Mar 2007.

[96] M W Young and S A Kay. Time zones: A comparative genetics of circadian clocks. *Nat Rev Genet*, 2(9):702–715, Sep 2001.

[97] W Yu, H Zheng, J H Houl, B Dauwalder, and P E Hardin. PER-dependent rhythms in CLK phosphorylation and E-box binding regulate circadian transcription. *Genes Dev*, 20(6):723–733, Mar 2006.

[98] H Zeng, P E Hardin, and M Rosbash. Constitutive overexpression of the *Drosophila period* protein inhibits *period* mRNA cycling. *EMBO J*, 13(15):3590–3598, Aug 1994.

[99] W F Zimmerman, C S Pittendrigh, and T Pavlidis. Temperature compensation of the circadian oscillation in *Drosophila pseudoobscura* and its entrainment by temperature cycles. *J Insect Physiol*, 14(5):669–684, May 1968.

[100] H Zouridis and V Hatzimanikatis. A model for protein translation: Polysome self-organization leads to maximum protein synthesis rates. *Biophys J*, 92(3):717–730, Feb 2007.